이 학 박 사 학 위 논 문

# Establishing strategies of genome assembly for unprecedented species

# 새로운 종에 대한 유전체 조립전략 수립

2017 년 2 월

서울대학교 대학원

생물정보협동과정 생물정보학전공

곽 우 리

# Abstract

## Establishing strategies of genome assembly for unprecedented species

Woori Kwak

Interdisciplinary Program in Bioinformatics

Seoul National University

Advancement of sequencing technologies provide more possibility to build unprecedented genomes. Building genome sequence is one of the most effective ways to identify the specific genomic features and evolutional history of unprecedented species. However, even though genome project can be conducted in more reasonable price than before, genome project still need quite amount of research funds. Therefore, many researchers who study non-model or novel organisms have difficulty in researches in genomic level. To solve this difficulty, this study mainly focused on not only building genome sequences of unprecedented species and but also practical use of genome assembly technique for resolving the limitations of previous methods. Custom

pipelines were built with various programs which have its unique characteristics for each purpose of study.

In chapter 1, the general background of NGS and *de novo* assembly was summarized. Characteristics of current sequencing technologies were presented, and the assembly algorithms and representative programs were listed.

In chapter 2, microbial genome project had been conducted. I built the complete genome of microbe using Pacbio RS II system. Using comparative genomic analysis, I identified potential genetic background of specific phenotype such as host immune system enhancement.

In chapter 3, abalone genome project had been conducted with various type of sequencing data. This is the first Haliotidae genome project and I identified specific evolutionary signature recorded in the abalone genome via comparative genomics analysis. And not only for abalone genome but also genomes and transcriptomes in abalone viscera were analyzed to identify the unknown origin of tumor suppression effect.

In chapter 4, I constructed effective complete mitochondrial genome assembly pipeline for unprecedented species. This pipeline includes various experimental process from different types of studies. Constructed pipeline was validated using real sequencing data and

mitochondrial genome sequence of *Menathais tuberosa* was successfully constructed.

Chapter 5 is about the microsatellite marker development pipeline for unprecedented species based on whole genome assembly. Constructed pipeline was tested for *Antheraea yamamai* genome and it successfully built microsatellite markers which showed possibility to be used for population genetics study.

Last chapter is about unaligned read assembly. Unaligned read assembly is one of the ways to solve the limitations of reference based re-sequencing study. Developed pipeline was applied to Hanwoo population data and it identified structural variation which might be related with its phenotypes.

Based on these studies, I expect my analysis and constructed pipelines can provide more understanding for unprecedented species and contribute to ameliorate limitations of current researches in this field.

**Key words**: genome assembly, mitochondrial genome, microsatellite, unaligned read,

**Student number** : 2014 - 30098

# Contents

# List of Tables

# List of Figures

# Chapter 1. Literature Review

## 1.1 Current sequencing technologies

### The beginning of sequencing.

Sequencing is defined as the process which decodes the nucleotide of DNA sequence of the genome. Even though Maxam and Gilbert developed the first modern sequencing technology in 1977(Maxam and Gilbert 1977), that of Sanger (Sanger et al. 1977) is known as the first generation sequencing method or the conventional sequencing method in these days. Sanger sequencing uses ddNTP (dideoxyribo nucleotides triphosphate) that do not have OH in 3' carbon of center sugar. The reason for using the ddNTP is for termination, and the oxygen in OH residue of 3' carbon provides the energy that can continue the chain reaction of DNA synthesis. However, ddNTP which do not have 3'-OH residue makes the chain reaction terminated, and Sanger sequencing methods are based on this characteristic of ddNTP. Using this termination mechanism, fragments of DNAs with one base pair length difference are amplified, and electrophoresis for ordering the DNA fragments is conducted. The nucleotide of the DNA can be identified following the order of each nucleotide. The early stage Sanger sequencing has short read length and small throughput of data generation. In 1986, Applied Biosystems(ABI) introduced automated DNA sequencing that uses fluorescent primer labeled differently for each ddNTP. The different fluorescent spectrum of each ddNTP is used in combined electrophoresis gel, and nucleotide ordering was conducted using a computer(Smith et al. 1986). Using such advanced method, one

can conduct sequencing more efficiently and quickly, compared to manual decoding. In 1995, the first-generation sequencing became truly automated with capillary electrophoresis.

**Rise of Next Generation Sequencing**

Sanger sequencing method is widely used and it contributed to the many types of research especially in bioinformatics field. However, the cost of the first-generation sequencing is expensive, and the amount of data generation is limited. To solve these limitations of Sanger sequencing, new sequencing technology named "Next Generation Sequencing" began to appear. This technology had characteristics such as low cost and rapid time for data generation compared to previous Sanger sequencing method, and it was quickly applied to various type of researches(Metzker 2009).

Pyrosequencing is known as the first commercialized NGS technology, and it was developed by Jonathan Rothberg(Rothberg and Leamon 2008). The core algorithm of this method is detecting the pyrophosphate(PPi) release on nucleotide incorporation. Released PPis are converted quantitatively to ATPs by ATP sulfurylase. Generated ATP provides energy for the luciferase-mediated conversion of luciferin to oxyluciferin. Oxyluciferin generates visible light which can be detected by camera and

intensity of light means the amount of synthesized nucleotide. In pyrosequencing, no different fluorescence is used for detection signal between 4 dNTPs. Therefore, each dNTPs—A, T, G, and C—are used once at a time, and then we figure out each base specific signal. In pyrosequencing, DNA synthesis is conducted until the end of homopolymer (repeats of same base sequence) at a time. In this case, detection of the number of synthesized bases relies on the amount of generated signal when homopolymer was elongated at a time. However, the intensity of the signal is not exactly proportional to the number of elongated bases in the real experiment because of various reason such as the limited efficiency of enzyme and interrupted signal sensing. This type of variation in detecting signal can affect the read length, and it resulted in different sequence length. For this reason, pyrosequencing generated InDel sequencing errors frequently, in homopolymer region compared to conventional Sanger sequencing. Pyrosequencing read the fragmented DNA sequence using single direction method, and paired-end read can be generated using the mate-pair library. Read length is 600 base pair in average in GS FLX system of Roche and it is almost close to the read length of Sanger capillary sequencing technology.

The most representative NGS platform is Illumina sequencing platform. This, represented by Hiseq, is the most popular and widely used

sequencing platform in 2016. Core algorithm of Illumina sequencing platform is SBS(sequencing by synthesis), and it is based on the nucleotide called reversible terminator. Reversible terminator blocks 3'-end for nucleotide incorporation similar to ddNTP used in conventional Sanger sequencing. However, as the name might suggest, a reversible terminator can recover its 3'-OH residue for elongation. Same as Sanger sequencing, Nucleotide which is with blocked 3'-OH is incorporated to the primer sequence, and the process of DNA synthesis is terminated. Each reversible terminator is labeled with a fluorescence dye, and a camera can detect it. After detecting elongated one base nucleotide, the 3'-end recover its OH residue. These three steps (nucleotide incorporation, detecting fluorescence, recover 3'-OH) consist one cycle which is core sequencing algorithm of Illumina sequencing platforms. Only one nucleotide can be detected in 1 cycle, so Illumina's sequencing platform is free from InDel type sequencing commonly occurred in pyrosequencing method. Illumina's sequencing platform read the fragmented DNA sequence using single or paired-end read method and read length vary from 50bp to 300bp. The read length of Illumina platforms is shorter than GS FLX system using pyrosequencing, but it is much cheaper to generate sequencing data. Illumina's sequencing

platform have lowest sequencing error rate among existing NGS platforms.

NGS technology is classified under two categories in these days, second and third generation technology. Two major sequencing platforms described above are classed as second generation sequencing. Sanger sequencing technology is categorized to first generation sequencing technology. Second generation technology has some characteristics compared to Sanger sequencing technology. First, the read length of second generation sequencing is shorter than Sanger sequencing. For example, the read length of Sanger sequencing is almost 1kbp, and it is much longer than Illumina's Hiseq. Second is the data generation throughput and time. Second generation sequencing generates high throughput sequencing data in short time compared to Sanger sequencing. For example, one Hiseq2500 device generates 10X coverage genome data of 20 peoples in almost one day, and it is tens of thousand times faster than Sanger sequencing device. Next is the low cost. It is expected that sequencing the genome of one individual will cost under $100 shortly. Forth, sequencing reactions are conducted in the miniaturized device compared to Sanger sequencing. Fifth, the error rate is higher than Sanger sequencing method. The error rate of Illumina platform and pyrosequencing is known as 0.26% and 1.07%. Moreover, this is greater than the error rate of Sanger sequencing method (about 0.1%)

**Advancement of Sequencing Technologies**

Second generation NGS platforms such as Illumina Hiseq and Roche GS FLX sequencing systems have some common limitations. First, these systems still use fluorescence detection system. Illumina system uses nucleotides labeled with the specific fluorescence color for each base. GS FLX system also detects the amount of light as a signal of the nucleotide incorporation. This type of technology must have imaging system for detecting the signal and errors can be generated and accumulated. For example, in each cycle of Illumina sequencing, fluorescence molecules in the cluster have to be removed for next nucleotide incorporation. However, sometimes all fluorescence molecules cannot be removed perfectly. Remained fluorescence molecules are accumulated and fluorescence signal is confused with the error signal. This is the reason for Illumina sequencing system has lower quality scores in end of the read. Imaging system using camera also can be a problem for the miniaturizing sequencer. In addition, the second generation sequencing relies on the PCR reaction for preparing the sequencing library. However, GC contents which mean the proportion of G and C nucleotide in total nucleotides can affect the PCR result and this is called “PCR bias”. All region of the whole genome cannot be amplified evenly, and high or low GC regions are more difficult to amplify using PCR. Therefore, sequencing results using the PCR-based

library are necessarily biased. Finally, the error rate of these sequencing technologies is still higher than conventional Sanger sequencing method. To resolve the limitations, researchers developed various sequencing platforms and technologies. Two commercialized sequencing platforms, ion torrent from life science and RS system from Pacific Biosystem, are representative examples.

Ion torrent(Rothberg et al. 2011) is based on the similar algorithm of GS FLX system which uses the byproduct of nucleotide incorporation like pyrophosphate. Instead of pyrophosphate, ion torrent system detects the hydrogen ion which is also a byproduct of nucleotide incorporation using PH value. This have benefits compared to pyrosequencing because fluorescence molecule and imaging device can be removed. With this benefits, sequencing process can be processed on a small semiconductor chip. So the sequencing device can be efficiently miniaturized. InDels error, the major error type of pyrosequencing, also can be reduced because the signal can be detected more accurately. In addition, sequencing speed is also increased because sequencing step is minimized. However, even though ion torrent has many benefits compared to the second generation NGS system, it is still based on the amplification of DNA fragment using emulsion PCR in library construction. Therefore,

sequencing result is not GC bias free, and DNA fragments with high or low GC ratio cannot be decoded efficiently.

Pacbio RS system(English et al. 2012) using the new sequencing algorithm. It fixed the DNA polymerase in the bottom of the well and the DNA synthesis process is conducted in that fixed point of bottom location. However, it still uses fluorescence molecules and imaging device like second generation sequencing, and it has same limitations like second generation sequencing technologies. However, even though pacbio RS system uses fluorescence system, it adopts single molecular sequencing technology. That means that RS system does not amplify the DNA fragment for library construction like other sequencing platforms (Illumina, Roche, and Ion torrent). Sequencing without PCR process has some benefits in GC bias and RNA-seq. Because the sequencing result contains more GC bias-free data, it can cover more genomic regions and transcriptome compared to another system. RS system also generates long read length sequencing data compared to other sequencing platforms. Long read length provide advantages in specificity, haplotype, and isoform. However, even though Pacbio RS system has many benefits, higher error rate (approximately >15%) is a big problem of RS system. The method called CCS system (circular consensus sequencing system)(Travers et al. 2010) has been developed for complementing this

weakness. Hairpin structure adaptor attaches to the end of DNA fragment, and the sequencing process can be repeatedly conducted for individual DNA fragment. DNA fragment is sequenced at least three times, and the consensus base call can efficiently reduce the error rate of RS system using independent sequencing reactions for the same location.

Illumina also makes up for the weak points. First, read length is improved. Miseq V2 which is the most recent sequencing platform of Illumina produces 300bp pair-end data. And almost 500bp of single read can be generated for metagenome community analysis using the overlapping library. Second, PCR-free library preparation kit provides unbiased sequencing library. PCR amplification and gel electrophoresis have been used for making a typical sequencing library for Illumina. PCR-free Library preparation(Kozarewa et al. 2009) kit does not conduct PCR and use magnetic bead base for DNA isolation in library preparation protocol. Because it does not amplify DNA using PCR, the genome coverage of sequencing in high or low GC contents region can be increased. So biased dispersion of sequencing coverage in genomic location is greatly reduced. This can be used to for various genomes with high AT regions. The third is the moleculo technology for long read data generation. This technology is developed base on the genome assembly research of Botryllus schlosseri(Voskoboynik et al. 2013). The main concept of this

technology is partitioning specific size of DNA fragment. It analyzes and conducts assembly for the partitioned genomic region based on the index sequences. This bring the same effect like genome size reducing, so molecule system can be useful for the high heterozygous genome.

Nanopore sequencing(Branton et al. 2008) is the third generation sequencing platform in the true sense of the word. It does not use fluorescence molecules and imaging devices. It does not amplify DNA fragments, and it conducts sequencing on a single molecule of DNA fragment. The prototype device of Oxford nanopore is very small (palm size), and it can conduct sequencing just connect USB 3.0 cable with a laptop computer. Algorithm of nanopore system is similar to Pacbio system. However, it identifies the nucleotide using the electronic signal from nanopore protein instead of fluorescence. Even though the accuracy and the throughput of Oxford nanopore have to be improved, this sequencing platform shows the blue print of future sequencing technology. Smaller sequencer, lower price, more high throughput, single molecule, etc.

In conclusion, the changes of sequencing paradigm are extremely fast. For example, GS FLX, the first-second generation NGS system, is excluded from the current market. In addition, generated data from different sequencers have different characteristics followed the

sequencing algorithms. For example, error profile of Pacbio system is totally different with second-generation sequencing platforms. Therefore, researchers who want to analysis NGS data have to understand the unique characteristics and principle of sequencing technologies for the selection of proper analysis method.

## 1.2 Genome assembly using NGS data

### Conventional genome assembly

Genome assembly indicates that rebuild the original sequence of DNA from fragment DNA sequence. This is essential to figure out the genome sequence of the organism because sequencing machine cannot read the whole genome sequence at one time. Genome assembly is often likened to jigsaw puzzle. Assembly process relies on the overlap region of two fragments of DNA sequence reads in the same string. Using this overlapped sequence, two sequence reads can be connected and extended to the longer sequence. Various programs like Celera Assembler(Myers et al. 2000), ARACHNE(Batzoglou et al. 2002), PCAP(Huang et al. 2003) were widely used for assembly of whole genome shotgun sequencing data. This algorithm is called OLC (Overlap Layout Consensus) method which is representatively used in Human genome project. Identifying the overlapping sequence information is needed for this assembly method. For this, all fragmented read has to be compared with another sequence one by one, and this makes the memory size of computing server demand increase proportionally with the number of input sequences. Therefore, this method is proper to read long with a short number of reads. However, whole genome sequence data using next generation sequencing consist of huge amounts of short reads and

comparing all the pair of short read sequences is almost impossible in these days computing system resources. Because of this reason, another assembly algorithm has to be applied for NGS data.

**Genome assembly using NGS data**

There are several characteristics of NGS data which make the genome assembly more complex. First, the length of generated reads is shorter or not that longer than conventional sequencing platforms. Even though the read length of current sequencing platforms continuously increased, it still needs many resources for conducting large genome assembly. Moreover, the read length of common NGS data from Illumina platforms is still shorter than Sanger sequencing and researchers attempts to overcome this limitation through high throughput of data generation. The second is huge amounts of data produced by NGS. This is the most distinguishable characteristics of NGS experiments which produce a massive amount data. However, this makes the required memory resources of server increase drastically. Most of NGS data assemblers manage these kinds of large sequencing data through the use of K-mers based assembly method. A K-mer means a series of the contiguous base of length K and K is a positive integer. Converting reads to the set of K-mers reduces the total amount of data efficiently. For example, if we

divide generated sequence read into 4-mer, there can be 256 combinations regardless of the read amount. Therefore, searching shared K-mers is easier than searching overlaps of each read from entire sequencing data. However, if positive integer K is small, the unique information of original reads is lost because the combination is limited. Moreover, opposite, if K value is large enough, the uniqueness of original read remained but the possible combination of divided K-mer is exponentially increased, and the amount of huge data from NGS platforms cannot be reduced efficiently. Third, the error rate of current NGS platforms is higher than conventional sequencing. This can induce assembly errors which make incorrect assembly or shortened contigs because assembly based on the K-mer rely on the perfect overlap of each pair. Therefore, many statistical approaches were developed to correct the errors of NGS data. Forth, repeat sequences make the assembly more difficult. Genome regions which share perfect repeats or the length of repeated sequence longer than read length of NGS data cannot be perfectly reconstructed using short reads. Because this is a technical problem, the only way to solve this is longer read length. The last one is the sequencing coverage of genomic regions is not uniform. GC bias is one of the reasons, and this can cause the poor assembly result.

Assemblers using NGS data are classified under 3 categories including greedy assemblers, overlap-layout-consensus assemblers, and de Bruijn graph assemblers. Even though the details of these three kinds of assemblers are different between each other, most programs use graphs based techniques. Greedy assemblers are based on the simple principle. It iteratively extends a read or contig by adding reads based on sequence overlaps. Iteration is continued until the read or contig cannot be extended more. What the choice of backbone reads for extension is based on the number of matching bases in the overlapping sequence. Assemblers using OLC (overlap-layout-consensus) method conduct the assembly in three stages. The first stage is overlap discovery. In the first stage, all pair-wise comparison of generated reads is conducted based on the precomputed K-mer contents. Based on the overlaps in the first stage, the overlap graph is built and optimized. The last stage is consensus stage based on multiple sequence alignment. De Bruijn graph assembler is the most widely used assembler for short read NGS data such as Illumina platforms. This method has the advantages of dealing with huge amount of sequencing data from NGS platforms. The De Bruijn graph is based on K-mers, and all pair-wise comparison of entire reads is unnecessary. Individual reads do not need to be stored, and redundant sequences are collapsed. This has benefits for handling a large amount of NGS data,

but this method is vulnerable to sequencing error. In OLC method, even though there are some sequencing errors in the overlap region, contig can be extended because OLC method did not require a perfect match. OLC method uses full read length so it can handle the little amount of sequencing errors for an extension. However, the specificity of read is greatly reduced following K value decreasing. So if assembler allows some errors in overlap, this can cause wrong assembly. Because of this, one of the most important thing for successful genome assembly using NGS data is error correction.

**The representative assemblers for NGS data**

SSAKE(Warren et al. 2007) is known as the first short read assembler, and it is designed to assemble single-end reads. This assembler is based on the greedy algorithm and index reads by their prefixes. Reads with minimum prefix overlaps are searched iteratively. SHARCGS(Dohm et al. 2007), VCAKE(Jeck et al. 2007) and QSRA(Bryant et al. 2009) use a similar greedy algorithm like SSAKE.

Newbler(Margulies et al. 2005) is a widely used OLC-based assembler specially designed for GS FLX sequence platform of Roche. It is usually used for small size genome like microbial genome. Newbler is an

exclusively designed assembler for GS FLX system, and it can use short read data from Illumina sequencer to support GS FLX sequencing data. However, GS FLX system is no longer used, so this assembler is not focused in these days.

CABOG(Miller et al. 2008) is the revised version of Celera Assembler originally designed for Sanger reads, and the pipeline is designed for 454 data. CABOG collapses the homopolymer to single bases to overcome the InDel error of pyrosequencing. However, it is also out of date with the service termination of GS FLX system.

Euler(Pevzner et al. 2001) is developed originally for Sanger reads and modified to operate on various data, 454 pyrosequencing reads(Chaisson et al. 2004), single-end Illumina reads(Chaisson and Pevzner 2008), and paired-end Illumina reads(Chaisson et al. 2009). It uses K-mer based assembly based on the Eulerian path algorithm. Like other K-mer based assembler, Euler also conducts error correction before building de Bruijn graph. Based on K-mer frequency distribution, it detects base-call errors from K-mers with low frequency. Because K-mers resulting from sequencing errors show much lower frequency while most true K-mers repeatedly appear many times. K-mers with frequency below threshold are removed or corrected based on the true K-mer sequence data. While error correction step before assembly is essential, this kind of

modification can mask true polymorphism or true K-mers which show low frequency by chance. When reads are converted to K-mers, there is some information lost compared to using whole reads directly. Read-threading step by laying entire read onto its graph can be a solution for recovering this lost information.

Velvet(Zerbino and Birney 2008) is a very popular assembler among de Bruijn assemblers. It uses graph simplification algorithm which collapses simple paths into single nodes, and this makes the graphs much simpler. Three parameters used in Velvet is known to affect the result of assembly. The first one is K value used for K-mers generation, which have to be an odd integer. Second is the minimum expected frequency threshold of K-mers for error correction. Finally, the expected genome coverage controls spurious connection breaking in the graph.

AllPaths(Butler et al. 2008) and Allpaths-LG(Gnerre et al. 2011) are also de Bruijn graph-based assemblers for large genome assembly. Allpaths begin with error correction preprocess similar to Euler's and Velvet. It uses the base quality score from sequencing machine for this process. Filtered reads may be retained if the substitution of two low-quality base makes its K-mers trusted or the read is essential for building a path between pair-end reads. AllPaths-LG focused on the improvements to the AllPaths algorithm like better error correction for remaining true

SNVs while filtering as many sequencing errors as possible, more efficient gap filling and scaffolding, and graph simplification which can show the more reliable result in the eukaryotic genome assembly. However, AllPaths-LG requires high memory resources compared to other assemblers and requiring at least one more specially designed library is weakness point of this assembler.

SOAPdenovo(Li et al. 2010) and SOAPdenovo2(Luo et al. 2012) are freely available large genome assemblers based on the de Bruijn graph. It has small memory resource requirement and especially SOAPdenovo2 uses sparse graph which can reduce more memory requirement while the results of assembly are maintained. SOAPdenovo can conduct contig assembly and scaffolding independently. Moreover, Gapcloser, the inner module of SOAPdenovo, shows good performance for gap filling process and it can be conducted independently with SOAPdenovo. SOAPdenovo is one of the most widely used programs for various genome projects.

Platanus(Kajitani et al. 2014) is specially designed for assembling high heterozygous genome. Assembling highly heterozygous diploid genome is challenging area because heterozygosity increases the complexity of the de Bruijn graph structure. It efficiently reduced bubble structures in the graph caused by heterozygosity of diploid genome using identity and

coverage of node. However, when I tested this assembler for my own genome project, it did not show good performance for high heterozygous diploid genome compared to other famous assemblers such as SOAPdenovo and IDBA. Therefore, this cannot be a golden standard for all heterozygous genome.

HGAP assembly algorithm(Chin et al. 2013) represented by SMRT Portal is an assembly algorithm used for long reads from Pacbio system. It is based on the OLC method and specially designed for error profile of Pacbio long reads. The error rate of Pacbio read is much higher(20%>) than Illumina sequencing system and genome assembly cannot be conducted properly without error correction. However, error profile of Pacbio system is different with Illumina sequencing system. Most of the errors are substitution errors in Illumina system, but InDel error is common in Pacbio system. Because of these different characteristics of Pacbio reads, genome rebuild needs more specific error correction and assembly algorithms. Constructed algorithm was validated via various microbial genome assembly project, and the paradigm of bacterial genome assembly moved to Pacbio system from Illumina and GS FLX system.

In addition to assemblers described above, there are a lot of genome assemblers with specific characteristics and these can be used for the

genome project. Among these assemblers, some assemblers showed good performance and these are world widely used. However, even though some famous assemblers show good genome assembly performance compared to others, there is no golden standard program for all genome. It is natural because each genome sequence has specific characters. Therefore, a best assembler for specific genome can be different and choosing assembler is one of the most important features for the successful genome project. This cannot be learned from text or lesson, and the only experience, trial-and-error can bring the best result for genome assembly.

**Process of genome assembly using NGS data**

*De novo* assembly using NGS data is conducted through 6 steps. First is raw data quality control. There is no golden standard for this, but here are some typically used conditions for raw data preprocess.

- Read with N bases more than 10%.
- Read with low quality bases more than half of the read length.(quality score < 20)

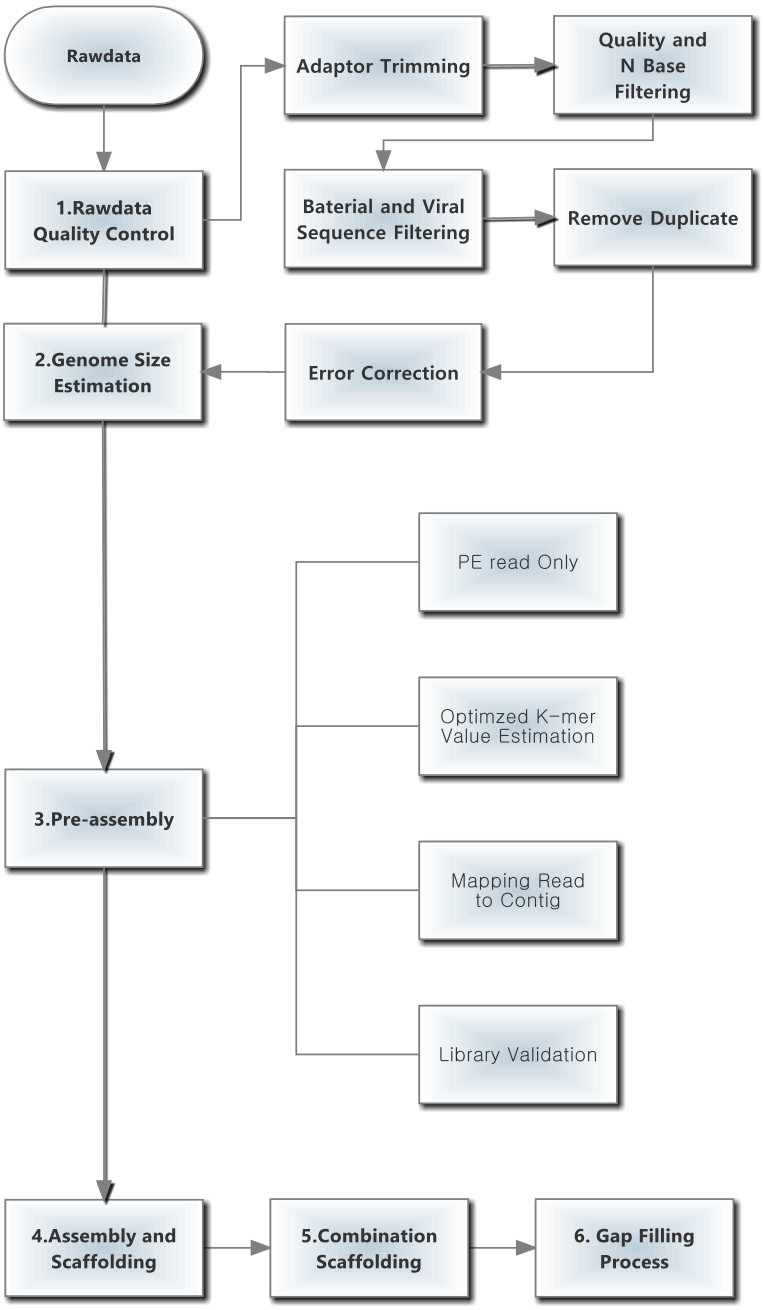



**Figure 1 - 1.** Overall *de novo* assembly Process

- Read contain more than 10bp adapter or primer sequence.
- Overlap paired-end read (except for Allpaths-LG).
- Two reads in each paired-end read are completely identical

After raw data quality control process, error correction process is conducted. Most of the assemblers have its own error correction module, but some independent error correction module like Quake(Kelley et al. 2010) and error correction module of AllPaths-LG are frequently used. The previous study showed that error correction module of AllPaths-LG has the best performance. Next, pre-assembly using only paired-end reads is conducted for checking the performance of assemblers and validating the mate-pair library. Mate-pair libraries of which the insert size is longer than 2kbp are mapped to the assembled scaffolds and insert size distribution is calculated. Based on the result of insert size distribution, researchers can check that the library was built properly. After initial contig assembly, stepwise scaffolding is conducted in order to get extended contig using the mate-pair library. Gaps in extended scaffolds are filled with paired-end and mate-pair library during a process called “gap filling”. Researchers repeatedly conduct the whole process described above with various algorithm, and the best result is chosen among the results.

# Chapter 2. Microbial Genome Assembly: Comparative analysis of the complete genome of *Lactobacillus plantarum* GB-LP2 and potential candidate genes for host immune system enhancement

## 2.1 Introduction

Respiratory infectious diseases caused by influenza viruses are increasingly becoming a global health concern (Ferguson et al. 2006). They have considerable negative effects on lifestyles and economies (Sander et al. 2009). Since the pandemics of avian-origin H5N1 and swine-origin H1N1 influenzas (Ferguson et al. 2004, Smith et al. 2009), researchers have begun to focus more on developing preventive agents for use until vaccines are developed. Many researchers have explored the potential use of probiotics, including *Lactobacillus* strains that exhibit antiviral activities, using different delivery methods in animals and humans (Boge et al. 2009, Leyer et al. 2009, Rizzardini et al. 2012, Kechaou et al. 2013, Park et al. 2013). Increasing evidence suggests that intranasal administration of *Lactobacillus* species could protect against influenza virus infection by modulating respiratory immune responses in animal models (Harata et al. 2011, Youn et al. 2012, Choi et al. 2015, Zelaya et al. 2015). A recent study found that both live and dead forms of *Lactobacillus rhamnosus* increase innate immune responses and attenuate pulmonary damage in respiratory tissues (Zelaya et al. 2015).

Previous research using genome sequencing and functional analyses has revealed the underlying mechanisms of how specific strains exhibit better performance in preventing and treating infectious intestinal

diseases via physical, biochemical, and immunomodulatory interactions (Lebeer et al. 2008, Liévin-Le Moal and Servin 2014). However, genomic evidence is not yet available to support their anti-infective activities related to the influenza virus. In this study, we sequenced the complete genome of *Lactobacillus plantarum* GB-LP2 isolated from the traditional Korean fermented vegetable. Previous research has demonstrated that it exhibits antiviral effects on influenza virus infection in mice following intranasal administration (Choi et al. 2015). Whole-genome assembly of the complete genome sequence of *L. plantarum* GB-LP2 was conducted and genomic contents were identified. A comparative phylogenetic tree revealed the evolutionary relationship between *L. plantarum* GB-LP2 and other previously reported *L. plantarum* strains. Additionally, a comparative genomic analysis was performed with seven other complete genome sequences of *L. plantarum* strains. I identified evolutionarily accelerated genes that might affect the phenotypic trait related to the anti-infective activities of influenza virus. This is the first report of a complete sequence and comparative analysis related to the anti-infective activities of influenza virus in *L. plantarum* species. The results will help clarify the anti-infective activities of probiotics and related genes.

## 2.2 Material and methods

### Strain Isolation and Whole-Genome Sequencing

Genomic DNA of the GB-LP2 strain was isolated and purified using an UltraClean Microbial DNA Isolation Kit (MoBio, Carlsbad, CA, USA) according to the manufacturer's protocol. The concentration and purity of the extracted DNA were determined using a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Approximately 5 μg of the extracted genomic DNA was sheared mechanically into 8–12 kb fragments using a Hydroshear system (Digilab, Marlborough, MA, USA). SMRTbell libraries were prepared for SMRT sequencing with C4 chemistry on a PacBio RS II system (Pacific Biosciences, Menlo Park, CA, USA). Libraries were purified using 0.45× AMPure XP beads to remove short inserts of <1.5 kb. The size distribution of the sheared DNA template was characterized using an Agilent 12000 DNA Kit (Applied Biosystems, Santa Clara, CA, USA), and concentration was determined using Invitrogen Qubit (Carlsbad, CA, USA). The sequencing primers were annealed to the templates at a final concentration of 5 nM template DNA, and DNA polymerase enzyme C4 was added according to the manufacturer's recommendations for small-scale libraries. A DNA/Polymerase Binding Kit P6 (Pacific Biosciences) was used to load the enzyme template-complexes and libraries onto

75,000 zero-mode waveguides (ZMWs). A DNA Sequencing Reagent 2.0 Kit (Pacific Biosciences) was used to sequence SMRT cells using a 120-min sequence capture protocol along with a stage start to maximize the subread length with PacBio RS II.

**Genome Assembly and Annotation**

Raw sequence data from the PacBio RS II system was filtered and assembled using the Pacbio SMRT portal system v2.3.0. The "RS_HGAP_assembly.3" algorithm (Chin et al. 2013) was employed and the genome size parameter was set to 3,300,000 bp using the Compute Minimum Seed Read Length option. Other parameters were set to default. Assembled contigs with a short contig length (<20,000 bp) and low coverage (<50X) were filtered for further analysis. To remove errors in the pre-assembled GB-LP2 genome sequence, an iterative polishing process was conducted until no genomic variants were identified. Genome annotation was completed using the RAST annotation system (Aziz et al. 2008) with default settings. COG annotation was conducted using WebMGA (Wu et al. 2011) and an annotation map was generated using DNA Plotter (Carver et al. 2009).

**Comparative Analysis**

For the comparative analysis, seven complete genomes and 24 draft genome sequences of *Lactobacillus plantarum* strains were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/genome/genomes/154). Average nucleotide identity (ANI) values were calculated for all 32 strains using JSpecies v1.2.1 (Richter and Rosselló-Móra 2009). An ortholog gene set for the eight complete genomes was built using the MESTORTHO method (Kim et al. 2008) and PRANK (Löytynoja and Goldman 2008) was employed for the multiple sequence alignment of each ortholog gene. After poorly aligned sites were eliminated using GBlocks (Talavera and Castresana 2007), orthologous sequences were concatenated to one sequence to construct a phylogenetic tree. MEGA6 (Tamura et al. 2013) was used to build the phylogenetic tree with the Neighbor Joining method, and bootstrap analysis was performed on the combined dataset sequences. The maximum likelihood method (codeml of PAML4) (Yang 2007) was used to estimate the dN (rate of non-synonymous substitution), dS ( rate of synonymous substitution), and evolutionarily accelerated genes under the branch and branch-site model.

**Strain Deposition and Complete Genome Sequence Accession Number**

The *L. plantarum* GB-LP2 strain was deposited in the Korea Agriculture Culture Collection (KACC, Korea) under number KACC 18511. The NCBI accession number of the complete genome sequence of this strain is SUB1096363.

## 2.3 Results

### General Features of *Lactobacillus Plantarum* GB-LP2

The *L. plantarum* GB-LP2 genome consists of a single circular DNA chromosome of 3,284,304 bp with a GC content of 44.57%. The GB-LP2 genome contains 3,250 open reading frames (ORFs), 76 tRNAs, and 16 rRNAs (Figure 2-1 and Table 2-1). Among the predicted ORFs, 2,435 genes (77.0%) were predicted as functional genes and 815 (33%) were unknown and hypothetical genes. Figure 2-2 presents the categorization of predicted ORFs based on the SEED subsystem categorization and COG functional categorization. Of the 2,435 genes, 1,922 ORFs were assigned to 25 SEED subsystem categories. One hundred twenty-four ORFs were assigned to the Cell Wall and Capsule category related to host antigen reaction; 34 ORFs were responsible for Capsular and extracellular polysaccharides; 55 ORFs were responsible for Cell Wall and Capsule; and 35 ORFs were responsible for Gram-Positive cell wall components. One hundred thirty ORFs were categorized into Cofactors, Vitamins, Prosthetic Group, and Pigments, and 11 ORFs were responsible for Thiamin Biosynthesis related to one of the main nutrient factors in the traditional Korean fermented vegetable. Fifty-one ORFs were assigned to Virulence, Disease, and Defense Subsystem including

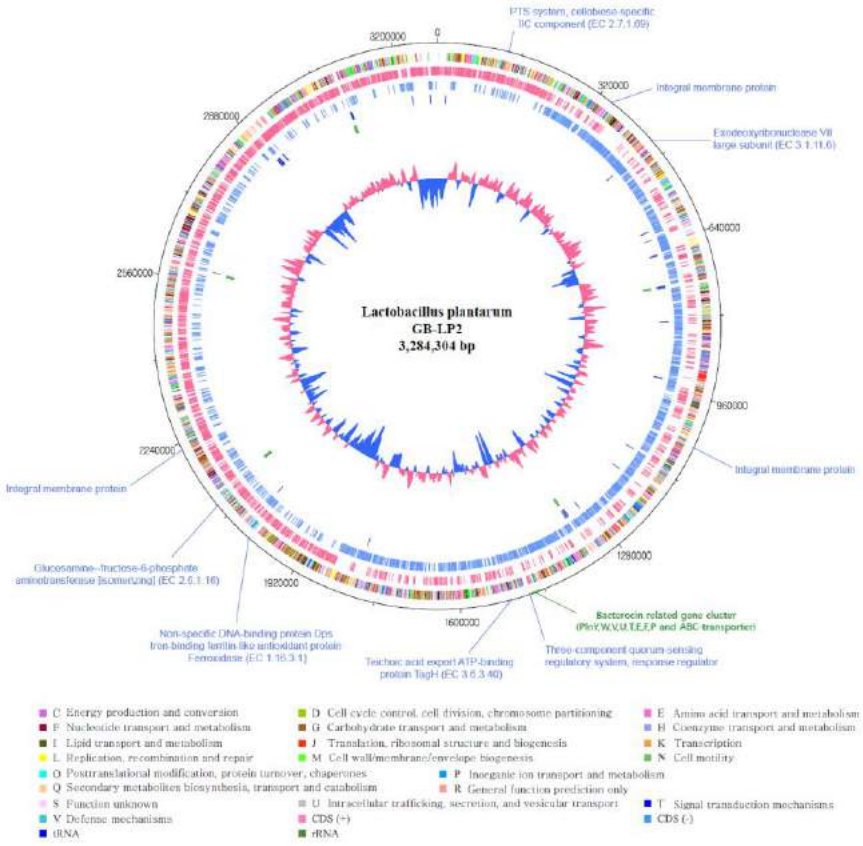



**Figure 2 - 1**. Genome map of *L. plantarum* GB-LP2.

Circles, from outer to inner, represent COG distribution, CDS in the leading strand, CDS in the lagging strand, tRNA, rRNA, and the GC content ratio. Functional genes are labeled around the outer circle as follows: evolutionarily accelerated genes in blue, genes related to antimicrobial activity in green.

**Table 2 - 1.** Comparison of the chromosomal properties of *L. plantarum* strains.

| Strain | GB-LP2 | P-8 | WCFS1 | ZJ316 |
|---|---|---|---|---|
| Genome size (bp) | 3,284,304 | 3,035,719 | 3,308,273 | 3,203,864 |
| GC content | 44.57 | 44.8 | 44.47 | 44.65 |
| Open reading frames | 3250 | 2926 | 3152 | 3110 |
| Average length (bp) | 844 | 849 | 884 | 864 |
| % of encoded genes | 77.0% | 77.2% | 78.0% | 72.6% |
| Annotated genes | 2435 | 2258 | 2457 | 2259 |
| Hypothetical and unknown proteins | 815 | 668 | 695 | 851 |
| tRNA | 76 | 70 | 74 | 63 |
| rRNA | 16 | 16 | 16 | 15 |
| ANI (average nucleotide identity) | 100% | 99.26% | 98.83% | 99.45% |

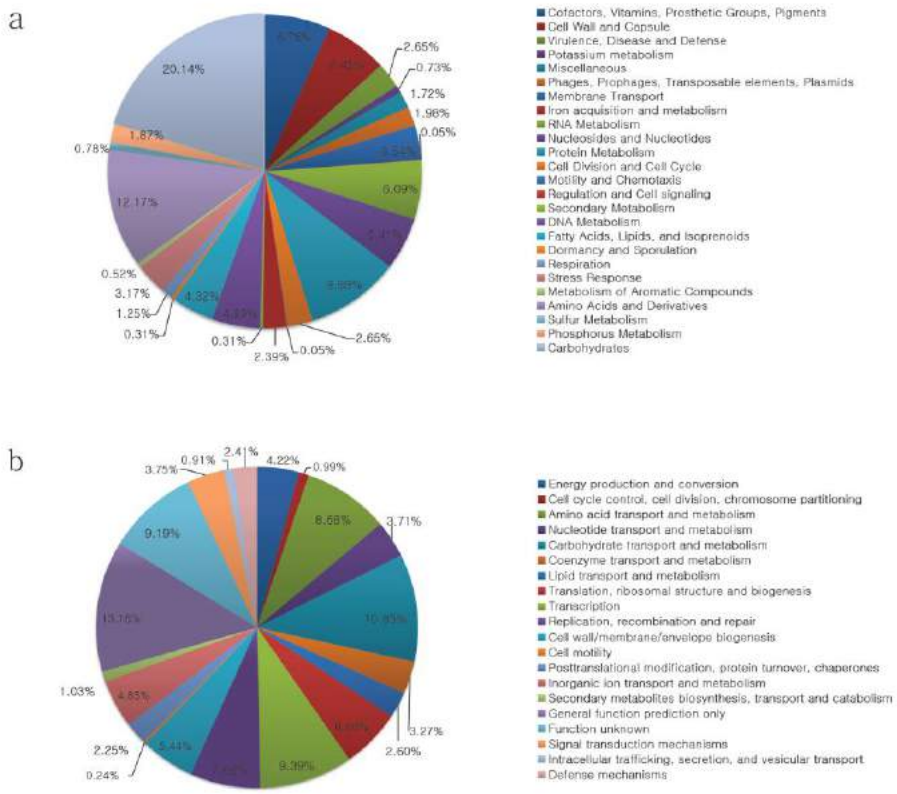



**Figure 2 - 2**. Functional categorization of all predicted ORFs in the *L. plantarum* GB-LP2 genome based on (a) SEED and (b) COG databases

resistance to antibiotics and toxic compounds such as bile, tetracycline, fluoroquinolones, and beta-lactam antibiotics. Based on the COG functional categorization, 2,535 ORFs (78.0% of all predicted ORFs) were classified into COG functional categories. Among these, 875 ORFs (40.8% of the COG-assigned ORFs) belonged to five major COG functional categories: 220 ORFs in category E (amino acid transport and metabolism), 275 ORFs in category G (carbohydrate transport and metabolism), 179 ORFs in category L (replication, recombination, and repair), 123 ORFs in category P (inorganic ion transport and metabolism), and 238 ORFs in category K (transcription).

**Comparative Tree Analysis**

Two ANI trees and one phylogenetic tree were constructed for a comparative tree analysis of the GB-LP2 strain. Two ANI trees were built with the 31 available genome sequences and seven available complete genome sequences in the NCBI database, respectively (Figures 2-3a and 2-3b). Six strains, including 16, P8, UCMA3037, TIFN101, IPLA88, and ZJ316, were grouped together with GB-LP2, and ZJ316 was regarded as the closest neighbor of GB-LP2 (99.45% of the ANI value). ZJ316 is a probiotic strain isolated from infant fecal samples; a previous study revealed that it has probiotic effects on pig growth and pork quality (Suo

et al. 2012). EDG-AQ4 and AY01 strains had 79.5% and 85.82% of the ANI value, respectively, and the results indicated that the genome sequences may differ even in the same microbial species. A phylogenetic tree using orthologous genes was generated for eight complete genome sequences using the neighbor joining method and bootstrapping 1000 times (Figure 2-3c). The bootstrap value for each node was 1.000, except for one node (0.992) between STIII and JDM1. The closest neighbor of GB-LP2 was ZJ316, as in the ANI tree, but the tree topology pattern of the phylogenetic tree did not exactly match that of the ANI tree. In the ANI tree, JDM1 was clustered with WCFS1 and B21. In the phylogenetic tree, however, JDM1 was clustered with STIII.

**Comparative Genomics using dN/dS Analysis**

To identify the evolutionarily accelerated genes in the GB-LP2 strain, dN/dS analysis, based on two models (branch and branch-site), was conducted for the orthologous gene set. dN/dS analysis based on the branch model revealed 10 evolutionarily accelerated genes (Table 2-2). Among the 10 evolutionarily accelerated genes, six (PTS system, cellobiose-specific IIC component (EC 2.7.1.69), Phosphomevalonate kinase (EC 2.7.4.2), Exodeoxyribonuclease VII large subunit (EC 3.1.11.6), Teichoic acid export ATP-binding protein TagH (EC 3.6.3.40),

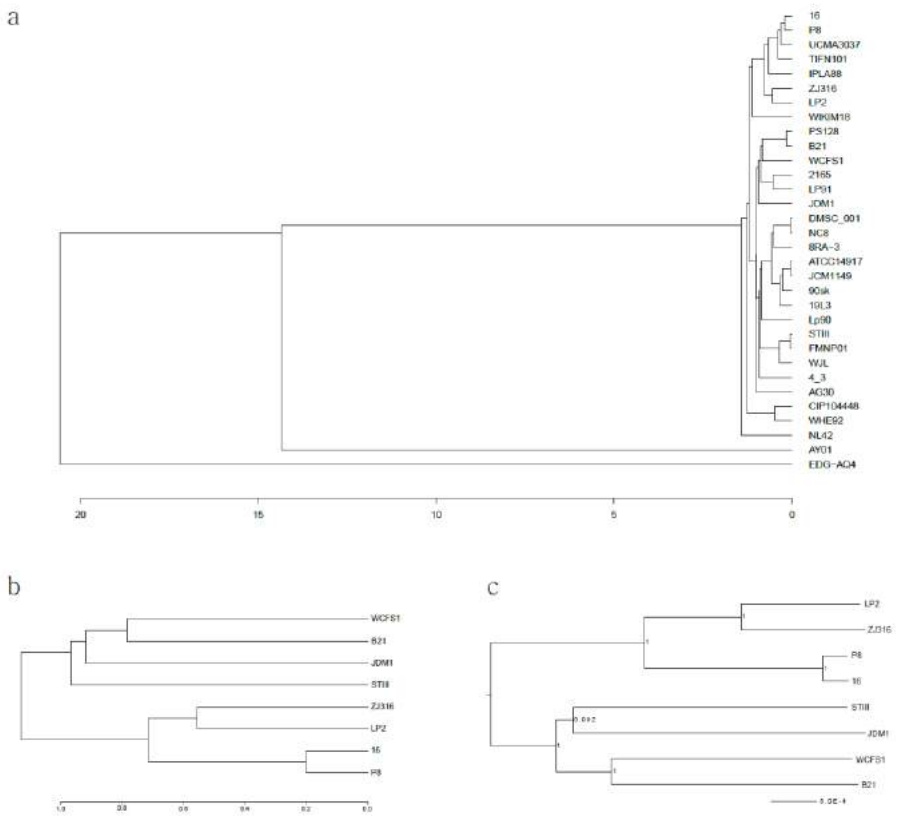



**Figure 2 - 3.** Comparative tree analysis**.**

(a) ANI tree analysis of 32 available genome sequences in the *L. plantarum* strain using Jspecies. (b) ANI tree analysis of GB-LP2 with seven available complete genome sequences of *L. plantarum.* (c) Phylogenetic tree analysis of GB-LP2 with seven available complete genome sequences.

Ribokinase (EC 2.7.1.15), and Glucosamine--fructose-6-phosphate aminotransferase [isomerizing] (EC 2.6.1.16)) were assigned an enzyme commission number. In the branch-site model, three genes were identified as evolutionarily accelerated genes (Table 3): two were integral membrane proteins and the other was the Non-specific DNA-binding protein Dps/Iron-binding ferritin-like antioxidant protein/Ferroxidase. Table 2-3 lists the differences in the amino acid sequences of the three genes. In the first integral membrane protein, 222a.a of GB-LP2 was changed to aspartic acid from glycine. Glycine is included in the non-polar amino acid group and aspartic acid is included in the polar group, as it has a negative charge. In the second integral membrane protein, 245a.a of GB-LP2 was changed to glutamine from lysine. Lysine is included in the positively charged polar group and glutamine is classified into the uncharged polar group. In genes related to ferroxidase, 256a.a, 290a.a, and 293a.a of GB-LP2 were changed to aspartic acid, valine, and threonine from asparagine, threonine, and valine, respectively. Valine is included in the non-polar amino acid group and threonine is included in the uncharged polar group.

**Table 2 - 2**. Evolutionarily accelerated genes identified in the branch model and related information for *L. plantaum* GB-LP2.

| Function | p-value | $\omega_{F.G.}$ | $\omega_{B.G.}$ |
|---|---|---|---|
| PTS system, cellobiose-specific IIC component (EC 2.7.1.69) | 1.14.E-04 | 0.5698 | 0.2039 |
| Phosphomevalonate kinase (EC 2.7.4.2) | 2.14.E-02 | 0.2974 | 0.0139 |
| Exodeoxyribonuclease VII large subunit (EC 3.1.11.6) | 4.62.E-02 | 0.8085 | 0.0833 |
| Integral membrane protein | 2.59.E-02 | 0.5161 | 0.0001 |
| Three-component quorum-sensing regulatory system, response regulator | 2.13.E-02 | 1.0444 | 0.0344 |
| Teichoic acid export ATP-binding protein TagH (EC 3.6.3.40) | 2.97.E-02 | 1.1167 | 0.0001 |
| Predicted transcriptional regulators | 1.26.E-02 | 0.4971 | 0.0001 |
| Ribokinase (EC 2.7.1.15) | 3.52.E-02 | 0.2069 | 0.0001 |
| Glucosamine--fructose-6-phosphate aminotransferase [isomerizing] (EC 2.6.1.16) | 1.76.E-02 | 0.2646 | 0.0001 |
| RNA methyltransferase, TrmA family | 2.63.E-02 | 0.2645 | 0.0001 |

**Table 2 - 3.** Evolutionarily accelerated genes identified in the branch-site model and related information for *L. plantaum* GB-LP2.

| **Function** | Integral membrane protein | Integral membrane protein | Non-specific DNA-binding protein Dps/ Iron-binding ferritin-like antioxidant protein/Ferroxidase | | |
|---|---|---|---|---|---|
| **Peptide length** | 256 | 320 | 294 | | |
| **p-value** | 2.72.E-03 | 4.75.E-02 | 2.66.E-05 | | |
| **$\omega 2_{F.G.}$** | 815.24 | 999.00 | 999.00 | | |
| **$\omega 2a_{B.G.}$** | 0.12 | 0.00 | 0.02 | | |
| **Proportion of 2a** | 0.004 | 0.004 | 0.008 | | |
| **Posterior probability** | 0.915 | 0.608 | 0.518 | 0.972* | 0.962* |
| **Amino acid sequence** | | | | | |
| **Position** | 222 | 245 | 256 | 290 | 293 |
| **LP2** | D | Q | D | V | T |
| **ZJ316** | G | K | N | T | V |
| **p-8** | G | K | N | T | V |
| **16** | G | K | N | T | V |
| **STIII** | G | K | N | T | V |
| **JDM1** | G | K | N | T | V |
| **WCFS1** | G | K | N | T | V |
| **B21** | G | K | N | T | V |

## 2.4 Discussion

### Modulation of innate immunity in the host

*L. plantarum* protects human intestinal cells from the invasion of pathogens via competitive adhesion, modulation of dendritic cells and T cells, and tight junction integrity (Liu et al. 2014). Additionally, the microdomain of the integral membrane protein (IMP) of *L. plantarum* plays a major role in the capacity of adhesion in intestinal epithelial cells, particularly in binding to the mannose receptor, which inhibits the Toll-like receptor (TLR) 5 pathway-mediated p38 MAPK signaling pathway after enteropathogenic *Escherichia coli* infection (Liu et al. 2012). I speculated that the evolutionarily accelerated genes of IMP in *L. plantarum* GB-LP2 that alter the binding capacity to receptors related to innate immunity could protect bronchial cells from the development of influenza virus infection and secondary infection, which may be associated with the preventive effect of GB-LP2 in mice challenged by influenza H1N1 (Choi et al. 2015). A teichoic acid export ATP-binding protein (TagH) (EC 3.6.3.40) is a subunit of the ABC transporter complex that exports teichoic acid, which is converted to wall teichoic acid (WTA) in the peptidoglycan layer and lipoteichoic acid (LTA) anchored to the cytoplasmic membrane. WTA and LTA bind to TLRs involved in the innate immune response (Bron et al. 2012), which may be associated with the capacity to

transport teichoic acid through the accelerated gene coded TagH-transporter complex. A glucosamine--fructose-6-phosphate aminotransferase (GlmS) (EC 2.6.1.16) has been evolutionarily accelerated in GB-LP2 compared with other *L. plantarum* strains. The GlmS is a rate-limiting enzyme in the biosynthesis of peptidoglycan, which constructs the mesh-like layer in bacterial cell walls and induces the host innate immune response (Barreteau et al. 2008, Bron et al. 2012). A change in the kinetics or modulation of the accelerated GlmS may be related to the capacity of peptidoglycan biosynthesis in GB-LP2.

**Survival of *L. plantarum* GB-LP2 in the respiratory tract prevents secondary infection**

The non-specific DNA-binding protein Dps plays a crucial role in protecting microorganisms from oxidative damage via mechanisms such as ferroxidase activity and DNA binding, increasing survival under harsh conditions such as starvation, and exposure to $H_2O_2$ (Chiancone and Ceci 2010). The evolutionarily accelerated gene Dps, if positively selected, may increase the survival of GB-LP2 in the respiratory tract, a nutrient-poor environment unlike the intestine, and prevent the host defense system against $H_2O_2$ produced by macrophages and neutrophils. Secondary infection caused by bacteria and influenza viruses increases the risk of morbidity and mortality (Joseph et al.

2013), and disruption of the balance between commensal bacteria and pathogens induces the development of infection on the bronchial epithelial cell barrier (Bosch et al. 2013). The evolutionary acceleration of Dps suggests that GB-LP2 survival may be related to the niche of overgrowth of pathogens in airways.

**Altering a three-component quorum-sensing regulatory system involved in plantaricin production**

The GB-LP2 chromosome contains all three genes related to a three-component quorum-sensing regulatory system (Maldonado-Barragán et al. 2009). An S-ribosylhomocysteine lyase (EC 4.4.1.31) produces autoinducer-2, a signaling molecule, to communicate among intra-species and inter-species to modulate population density and gene expression. The location of a histidine protein kinase gene and a response regulator gene in the cluster of bacteriocins of *L. plantarum* (plantaricins) indicates the quorum sensing-mediated regulation of plantaricin production (Sturme et al. 2007). Interestingly, a response regulator gene of GB-LP2 has been evolutionarily accelerated in an environment of vegetable-based fermentation (i.e., solid surface fermentation), which may increase the capacity of adhesion to the surface of the respiratory tract and competition with pathogens in the enhanced

quorum sensing-mediated regulation response compared with other *L. plantarum* strains (Maldonado-Barragán et al. 2009).

**Adaptation to the plant environment**

A comparative genomic analysis detected GB-LP2-specific genes such as the cellobiose-specific IIC component in the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS) (EC 2.7.1.69). The PTS is a major carbohydrate active-transport system in bacteria that catalyzes the phosphorylation of sugar substrates to cross the microbial cell membrane. *L. plantarum* GB-LP2 has a distinguished cellobiose-specific IIC component rather than the lactose-specific form, indicating the evolutionarily accelerated adaptation of GB-LP2 to vegetable-based fermentation.

In conclusion, this comparative analysis yielded potential candidate genes of *L. plantarum* GB-LP2 that may prevent infection and mortality from the influenza virus among humans via evolutionarily accelerated genes related to the innate immune response, survival, and quorum-sensing regulatory systems in the respiratory tract. Further research focusing on these evolutionarily accelerated genes will help clarify their preventative mechanisms against influenza virus and secondary infection in airways, and may help prevent seasonal flu and epidemic respiratory virus infections.

# Chapter 3. Eukaryotic Genome Assembly: Pacific abalone genomes provide insight into Haliotidae evolution and potential driver of tumor suppression effect

## 3.1 Introduction

Abalones are large marine snails in the family Haliotidae and the genus Haliotis belonging to the class Gastropoda of the phylum Mollusca. Family Haliotidae contains only one genus, Haliotis, and this single genus is known to contain several species of abalone. With 18 additional subspecies, the most comprehensive treatment of the Haliotidae considers 56 species valid(Appeltans et al. 2012).

Abalone is a widely used ingredient in East Asian cuisine, and it is a valuable food resource due to the richness in protein and other nutrients(Elliott 2000, Gordon and Cook 2004). The main body part used for food is its muscular foot, and it occupies most of the whole body. Even though the enlarged foot is a common phenotype between Gastropoda species, foot of abalone is more particular. Muscular foot of abalone is known to have the strongest attaching and grasping power among all gastropods, and historical record related to this particular phenotype also remained. According to the Annals of the Joseon Dynasty of Korea (in 1790), the King excluded the abalones, which are strongly attached to the rocks in the bottom of the sea, from donation to the palace, because many female divers were sacrificed during the abalone harvest. As well as enlarged muscular foot, the abalones are utilized as a research resource for ecological and evolutionary studies. Because they usually inhabit

the place where they first settled, abalone habitation impacts algal communications connected with the reef ecosystem(Hamer and Victoria 2010). And they are also utilized as models for evolutionary divergence, recognized for their variety in phenotypic appearance and global distribution(Estes et al. 2005, Panhuis et al. 2006). In addition, its role of medicine for tumor is recently focused. Previous studies showed that the visceral extract of abalone has tumor suppression effect on human breast and lung cancer(Lee et al. 2010, Suleria et al. 2015).

While the characteristics of abalone seem imperative, genomic research on the species is unsatisfactory, and the need for high-quality genomic information on abalone has increased to understand the genetic background of these characteristics. So far, linkage maps with microsatellite markers(Sekino and Hara 2007) or amplified fragment length polymorphisms(Liu et al. 2006) were constructed and had served as an efficient tool to achieve extensive genome coverage. However, the vast majority of previous research focused on the fertilization(Lee et al. 1995, Swanson and Vacquier 1998, Yang et al. 2000), shell formation(Menig et al. 2000, Lin and Meyers 2005), species classification(Kim et al. 2000, Coleman and Vacquier 2002), or economic traits(Kawamura et al. 1995, Kawamura and Takami 1995, Mai et al. 1995,

Viana et al. 1996); an in-depth analysis of the whole genome of abalone is yet to be conducted.

In this study, I present the first Haliotidae genome of *Haliotis discus hannai*. The goal is to uncover the history encrypted in abalone genome with a series of analyses. Biggest genome size in the class and genetic construction of abalone specific phenotypes is investigated by comparative genomics. The phylogeny, estimated divergence time and demographic history for of 4 pacific abalone species were checked. A population admixture analysis on Pacific abalone species around South Korea, via Genotyping-by-Sequencing (GBS), was also conducted to identify the underlying population structure as a research resource in an ecological study. Additionally, to gain insight into the molecular background behind the anti-cancer effect of visceral extract, I employed community and meta-transcriptome analysis. This work is the first large-scale genome analysis conducted for the species in Haliotidae family, and I expect our study settle some limitations of previous studies.

## 3.2 Materials and Methods

### A. ORGANISM BACKGROUND

#### A1. Taxonomy

Abalone are large marine snails in the family Haliotidae and the genus Haliotis belonging to the class Gastropoda of the phylum Mollusca. Family Haliotidae contains only one genus, Haliotis, and this single genus is known to contain several species of abalone. With 18 additional subspecies, the most comprehensive treatment of the Haliotidae considers 56 species valid(Appeltans et al. 2012)**.** Out of them, four economically valuable Pacific abalone species consist of *H. gigantea, H. madaka, H. discus discus*, and *H. discus hannai*.

#### A2. Habitation

Abalones are one of the most important marine gastropod molluscs that inhabit various coastal regions of the world. They live in near low tide lines or reefs that are seaweed-rich and provide pure water conditions. It is well known that abalone habitation impacts algal communications connected with the reef ecosystem, so they are often utilized for ecological research(Hamer and Victoria 2010). Moreover, their body size is related to their habitations; larger

abalones live in temperate water but smaller abalones live in tropical water. In evolutionary divergence, they are frequently utilized as model organisms, recognized for their variety in phenotypic appearance and global distribution(Estes et al. 2005, Panhuis et al. 2006).

**A3. Morphology**

As one of marine snails, abalone has a notably big foot. Its foot is a popular food source, which is densely covered with cilia that accounts for movement and adhesion using van der waals force (Lin et al. 2009). Abalones have layered shells of “brick-like” mantles, which are similar to other marine animals (Lin and Meyers 2005). The shells have open respiratory pores near the outer most layers that guides in/outflow of water to and from their branchial chamber. The magnitude of this water flow has been reported to be controlled by the lateral cilia of the gill lamella (Taylor and Ragg 2005). Abalone shell also includes a thick inner layer constructed by nacre (mother-of-pearl), which in many species is highly iridescent; this gives rise to a range of strong alternating colors that make the shells attractive to humans as decorative objects, jewelry, and as a source of colorful mother-of-pearl.

## A4. Reproduction and growth

Abalones are gonochoristic. They broadcast spawn, releasing their reproductive products into the ocean. Fertilization occurs in ocean water where the fertilized egg (zygote) develops to trochophora, a typical gastropod larval form. The trochophora are spherical shaped, and has a few ciliary bands that motors the body for free-swimming. These bands later develop into velum, a ciliated semi-circular structure resembling fins or wings, when trochophora are morphed to veliger. The velum is the main means of propulsion and food collection in this stage. From the rear of the velum, shell is secreted and covers the whole dorsal surface; hereby, the larvae go through a plankton stage until velum is degraded and falls to the bottom of the sea to morph into its adult form

## A5. Usages in food and medicine

Abalones are widely recognized as healthy and well-being foods. They are good source of proteins, lipids, essential fatty and amino acids (Latuihamallo and Apituley 2015), and they are rich in vitamin B1-2, taurine, and minerals including calcium and phosphorus (De Zoysa 2013, Lim 2014). Traditionally in Korea, it is believed that abalone could restore liver, kidney, and stomach

health. Recently, it is in the limelight that abalone has various biological activities involved in therapeutic potential: anti-oxidant, anti-thrombotic, anti-inflammatory, anti-microbial and anti-cancer (Suleria et al. 2015). Especially, abalone visceral extract display cytotoxic and antioxidant activities (Lim 2014), and has been demonstrated of its abilities to suppress tumor progression associated with Cox-2 level modulation and CD8+ T cell activity (Lee et al. 2010).

**A6. Economic values**

Abalones are also important fishery and aquaculture animal which is one of the highly prized sea food delicacies. Total global supply of abalone has increased fivefold since 1970's. In order to prevent indiscreetly fishing abalones, legal landings from abalone fisheries have made fishery productions decreased gradually from 19,720 mt to 7,486 mt, but have made farm productions increase explosively from 50 mt to 103,464 mt in the past forty years (Cook 2014). Therefore, the demand for expansion of knowledge in genetic characters related to various phenotypes of abalones increased for the breeding. Former researches, however, mainly dealt with developments on farming technologies and of fodder of abalones (LEE et al. 2001, Moon et al.

2006, Yoo and Chung 2007), or focused on food processing (Shin et al. 2008, Koh et al. 2009, Jang et al. 2012, Lim et al. 2013).

## B. GENOME SEQUENCING AND ASSEMBLY

### B1. Sample Collection for Draft Genome Assembly

A *Haliotis discus hannai* sample was collected from the brood stock at the Genetic and Breeding Research Center (GBRC) of the National Fisheries Research & Development Institute (NFRDI) on Geoje Island, Korea. Hemolymph was withdrawn from the sole side foot muscle using a syringe.

### B2. Library Construction and Data Generation

For genomic DNA extraction, hemocytes were harvested from fresh hemolymph by centrifugation at 3000 × rpm for 5 min at 4℃. Genomic DNA was extracted using a DNeasy Animal Mini Kit (Qiagen, Hilden, Germany). I used multiple sequencing platforms (Illumina Hiseq2000, Nextseq500 and Pacbio RS II) for whole genome sequencing. All sequencing processes were performed following manufacturer instructions. Two paired-end libraries and five mate-pair libraries were constructed. Two paired-end libraries were constructed with insert sizes of about 250bp and 350bp. Insert sizes of about 3k, 5k, 8k, and 10k libraries were constructed using Illumina Nextera mate-pair library construction protocol, and 20kb library was constructed for Pacbio

RS II system. Detailed information of constructed library and generated sequencing data is listed in Table 3-1.

### B3. K-mer Distribution and Genome Size Estimation

K-mer distribution of paired-end library provide valuable information about the target genome. First, quality control process of generated raw data was conducted for downstream analysis. Quality of raw data was checked with FASTQC and adapter sequences were removed via Trimmomatic(Bolger et al. 2014), for paired-end libraries, and Nxtrim(O'Connell et al. 2015), for mate-pair libraries. K-mer frequency analysis of the abalone genome was conducted using paired-end library with 350bp insert-size and the jellyfish(Marçais and Kingsford 2011) command-line program. As a result, 19-mer distribution of *Haliotis discus hannai* genome was generated (Figure 3-1). Genome size estimation based on the 19-mer distribution was conducted through "Estimate genome size.pl" code (https://github.com/josephryan/estimate_genome_size.pl/ wiki/Estimate-genome-size.pl". Estimated genome size of *H.discus hannai* using 19-mer distribution was about 1.65Gb.

**Table 3 - 1.** Summary statistics of generated whole genome shotgun sequencing data.

| Library Name | Library Type | Insert Size | Platform | Read Length | No. Read | Total bp |
|---|---|---|---|---|---|---|
| 250bp | Paired-end | 250 | Nextseq500 | 150 | 876,529,480 | 131,440,418,087 |
| 350bp | Paired-end | 350 | Hiseq2000 | 101 | 1,413,620,786 | 142,775,699,386 |
| 3k | Mate-pair | 3,000 | Nextseq500 | 150 | 580,064,464 | 85,689,154,056 |
| 5k | Mate-pair | 5,000 | Nextseq500 | 150 | 468,432,888 | 69,966,139,205 |
| 8k | Mate-pair | 8,000 | Nextseq500 | 150 | 335,132,792 | 50,109,845,012 |
| 10k | Mate-pair | 10,000 | Nextseq500 | 150 | 569,376,096 | 85,080,237,236 |
| 20k | P6-C4 | 20,000 | Pacbio RS II | 10,094 (average) | 1,573,020 | 15,879,626,978 |
| **Total** | | | | | | 580,941,119,960 |

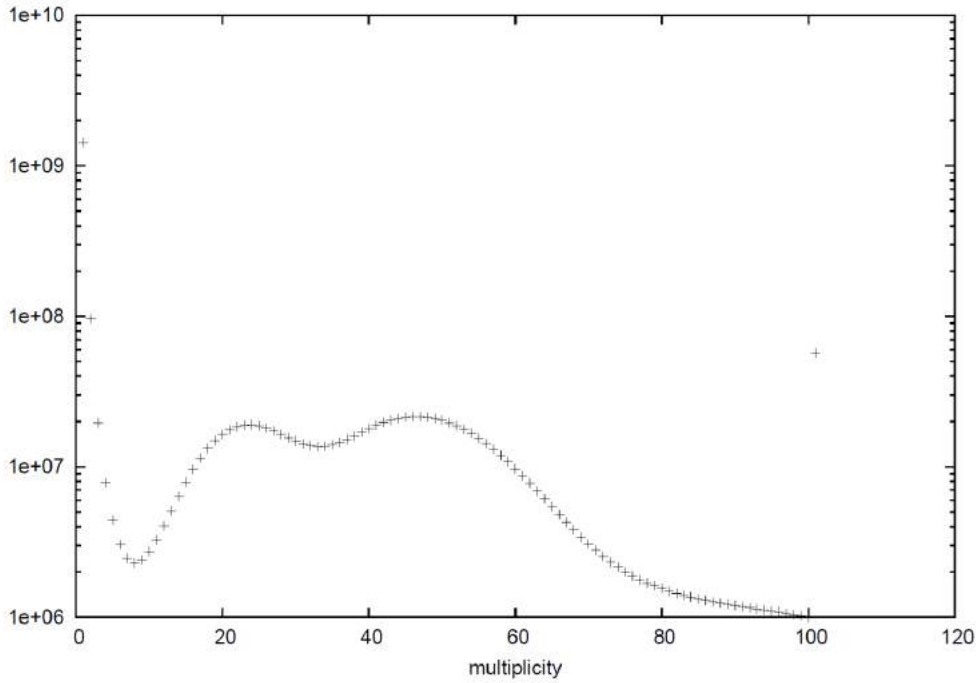



**Figure 3 - 1.** 19-mer distribution of using jellyfish with 350bp paired-end whole genome sequencing data.

**B4. Genome Assembly**

Based on the 19-mer distribution of paired-end reads, there was a second peak located in the half x-axis of the main peak. This result indicated that *H.discus hannai* genome had high heterozygous genetic character and probable DNA contamination of other organism. Before genome assembly, raw reads from Hiseq2000, Nextseq500 paired-end and mate pair were therefore pre-processed by bacterial sequences, duplicates, and ambiguous nucleotides. The resulting high-quality sequences were used in subsequent assembly. Error correction and contig assembly was conducted using clc_assembler from the CLC Assembly Cell (http://www.clcbio.com/products/clc-assembly-cell/) software pipeline, which is said optimal, to present the best *de novo* results, compared to a variety of assemblers with paired-end reads. And then scaffolds were built with the mate-pairs and Pacbio RS II reads sequentially by SSPACE(Boetzer et al. 2011) and PBJelly2(English et al. 2012). Finally, along with any missing reads, and reassembled using Gapcloser(Luo et al. 2012) in an attempt to close the gaps within each scaffold with -l 155 and -p 31, iteratively. Summary statistics for final assembly is given in Table 3-2.

**Table 3 - 2.** Summary statistics for the *Haliotis discus hannai* draft genome.

| **Assembled Genome** | |
|---|---|
| Size(1n) | 1.86 Gb |
| GC level | 40.47% |
| No. scaffolds | 80,032 |
| N50 of scaffolds (bp) | 200,099 |
| N bases in scaffolds (%) | 116 Mb (6.25%) |
| Longest(shortest) scaffolds (bp) | 2,207,537 (854) |
| Average scaffold Length (bp) | 23,309.12 |

**B5. Assembly Quality Evaluation**

Read remapping and CEGMA (Core Eukaryotic Genes Mapping Approach)(Parra et al. 2007) and BUSCO (Benchmarking Universal Single-Copy Orthologs)(Simão et al. 2015) were employed for quality assessment of *H.discus hannai* draft genome. 248 core eukaryotic genes were used in CEGMA analysis, andmetazoan dataset for BUSCO analysis. The results of each method is shown in Table 3-3, 3-4, 3-5.

**Table 3 - 3.** Summary of whole genome read mapping using paired-end read with Bowtie2.

| Categories | Alignment Rate |
|---|---|
| Total number of reads pairs | 423,350,776 (100%) |
| Concordantly 1 time | 177,892,969 (42.02%) |
| Concordantly > 1 time | 199,497,044 (47.12%) |
| Discordantly 1 time | 3,307,932 (0.78%) |
| 1 time in mixed mode (singleton) | 14,886,969 (3.52%) |
| > 1 time in mixed mode (singleton) | 27,135,765 (6.41%) |
| Overall mapping rate including singletons | 94.89% |

**Table 3 - 4.** Summary statistics of the CEGMA analysis result based on 248 CEGs for *Haliotis discus hannai* genome.

| | #Prots | %Completeness | #Total | Average | %Ortho |
|---|---|---|---|---|---|
| Complete | 95 | 38.31 | 111 | 1.17 | 14.74 |
| Partial | 198 | 79.84 | 291 | 1.47 | 36.87 |

.

# Prots = number of 248 ultra-conserved CEGs present in genome

# %Completeness = percentage of 248 ultra-conserved CEGs present

# Total = total number of CEGs present including putative orthologs

# Average = average number of orthologs per CEG

# %Ortho = percentage of detected CEGS that have more than 1 ortholog

# CEG : core eukaryotic genes

**Table 3 - 5.** Summary statistics of Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis for *Haliotis discus hannai* genome based on Metazoans DB.

| Categories | #Genes | Percentage |
|---|---|---|
| Complete Single-Copy BUSCOs | 609 | 72.2% |
| Complete Duplicate BUSCOs | 48 | 5.7% |
| Fragmented BUSCOs | 130 | 15.4% |
| Missing BUSCOs | 104 | 12.3% |

## C. REPEAT IDENTIFICATION AND GENE ANNOTATION

### C1. Repeat Element Identification

For gene prediction of the assembled sequence and comparison of repetitive elements with *Lottia gigantea* (Limpet), the closest neighbor of *H.discus hannai* available in Ensembl database, sequence patterns (i.e. repeats) were identified using RepeatMasker(Tarailo-Graovac and Chen 2009) with Repbase(Jurka et al. 2005). RepeatModeler, which includes RECON(Bao and Eddy 2002), RepeatScout(Price et al. 2005) and TRF(Benson 1999), was used to build custom database of *H.discus hannai* and *L.gigantea.* After custom library construction, RepeatMasker with RMBlast was used for each genome with 'no_is' option, using repeat libraries from

RepeatModeler and Repbase. Identified repeat elements were parsed for further analysis using a perl code named "One code to find them all"(Bailly-Bechet et al. 2014).

### C2. Sample Collection and Data Generation for Transcriptome of *H.discus hannai*

Six *H. discus hannai* organ tissues; hemocytes, digestive duct, gill, hepatopancreas, mantle, and ovary; were collected for gene expression

profiling and supporting *ab initio* gene prediction. To obtain high-throughput transcriptome data, I performed Illumina-based NGS sequencing. Total RNA was then quantitated using Nanodrop spectrophotometer (Thermo Scientific) and quality-assessed by RNA 6000 Nano assay kit (Agilent) and Bioanalyser2100 (Agilent). Library construction for all prepared samples was conducted using Illumina Truseq library preparation kit with manufacturer's protocol. NGS sequencing libraries were generated from one microgram of total RNA using Truseq RNA Sample Prep Kit (Illumina) according to the manufacturer's protocol. Poly-A containing RNA molecules were purified using poly-T oligo attached magnetic beads. After purification, the total poly $A^+$RNA was fragmented into small pieces using divalent cations under elevated temperature. The cleaved mRNA fragments were reverse transcribed into first strand cDNA using random primers. Short fragments were purified with a QiaQuick PCR extraction kit and resolved with EB buffer for end reparation and addition of poly (A). Subsequently, the short fragments were connected with sequencing adapters; each library was separated by adjoining distinct MID tags. The resulting cDNA libraries were then paired-end sequenced (2x101bp) with HiSeq2000 system. Detailed information about the generated transcriptome data is summarized in Table 3-6.

For more accurate gene prediction of *H.discus hannai* genome, I also employed Iso-seq, using Pacbio RS II system, for full length transcript sequencing. Seven organ tissues; mantle, gills, hepatopancreas, muscle, gonad, intestine, and hemocyte; were collected. Library construction and sequencing process was conducted following the manufacturer instruction. Detail information of generated and filtered transcriptome data using Iso-seq is summarized in Table 3-7 and 3-8.

**C3. Gene Prediction and Consensus Gene set Build**

Genes were predicted through three different algorithms: *ab initio*, RNA-seq transcript based, and protein homology-based. Overall workflow of consensus gene set build is in Figure 3-2.

For RNA-seq transcript based prediction, transcriptome data from six organ tissues were aligned to the assembled genome sequence using Tophat(Trapnell et al. 2009), and transcript structure was predicted through Cufflinks(Trapnell et al. 2012). Summary statistics of Tophat alignment and cufflinks is shown in Table 3-9 and 3-10.

Iso-seq data, RNA-seq using Pacbio RS II system, was also used to build accurate gene set for *Haliotis discus hannai*. Filtered iso-seq reads were aligned to draft genome using PASA pipeline (Table 3-11).

The homology-based method employs complete protein sequences from diverse taxonomical genomes, which is fit to our model. For *Haliotis discus hannai,* the following 8 species were utilized: *Lottia gigantea, Crssostrea gigas, Aplysia california, Strongylocentrtus purpuratus, Branchiostoma floridae, Danio rerio, Oncorhynchus mykiss and Homo sapiens*. Those protein sequences were aligned to *Haliotis discus hannai* genome using TBASTN (E-value ≤ 1E-4)(Altschul et al. 1997). Then the homologous genome sequences were aligned to the matched proteins using Exonerate(Slater and Birney 2005) to predict the accurate spliced alignments. Table 3-12 summarizes the alignment results of known proteins in various species.

For *ab initio* gene prediction, Augustus(Stanke et al. 2008) was trained using RNA-seq data and known Protein, by using the complete transcriptome as training matrix for HMM. Fgenesh(Solovyev et al. 2006) and Geneid(Blanco et al. 2007) were also used. Used parameters and number of predicted genes is given in the Table 3-13.

**Table 3 - 6.** Summary statistics of generated transcriptome data for six organ tissue using Illumina platform.

| Library Name | Library Type | Platform | Read Length | No. Read | Total bp |
|---|---|---|---|---|---|
| Blood | Paired-end | Hiseq2000 | 101 | 53,525,950 | 5,406,120,950 |
| Digestive duct | Paired-end | Hiseq2000 | 101 | 56,485,666 | 5,705,052,266 |
| Gill | Paired-end | Hiseq2000 | 101 | 66,415,882 | 6,708,004,082 |
| Hepatopancreas | Paired-end | Hiseq2000 | 101 | 58,467,176 | 5,905,184,776 |
| Mantle | Paired-end | Hiseq2000 | 101 | 65,741,776 | 6,639,919,376 |
| Ovary | Paired-end | Hiseq2000 | 101 | 60,997,100 | 6,160,707,100 |
| Total | | | | | 36,524,988,550 |

**Table 3 - 7.** Summary statistics of generated transcriptome data using Pacbio Iso-seq protocol.

| | **No. cell** | **Total base** | **bp/cell** | **No. transcripts** | **No. High quality transcripts** |
|---|---|---|---|---|---|
| *Haliotis discus hannai* | 8 | 9,865,038,046 | 1,257,170,610 | 72,219 | 40,993 (56.8%) |

**Table 3 - 8.** Summary statistics of filtered transcripts data using Pacbio Iso-seq.

| | **No. transcripts** | **N50** | **Min, bp** | **Mean, bp** | **Max, bp** | **Total, bp** | **N%** | **GC%** |
|---|---|---|---|---|---|---|---|---|
| *Haliotis discus hannai* | 40,993 | 2,457 | 316 | 2,419.52 | 8,573 | 99,183,421 | 0 | 43.34 |

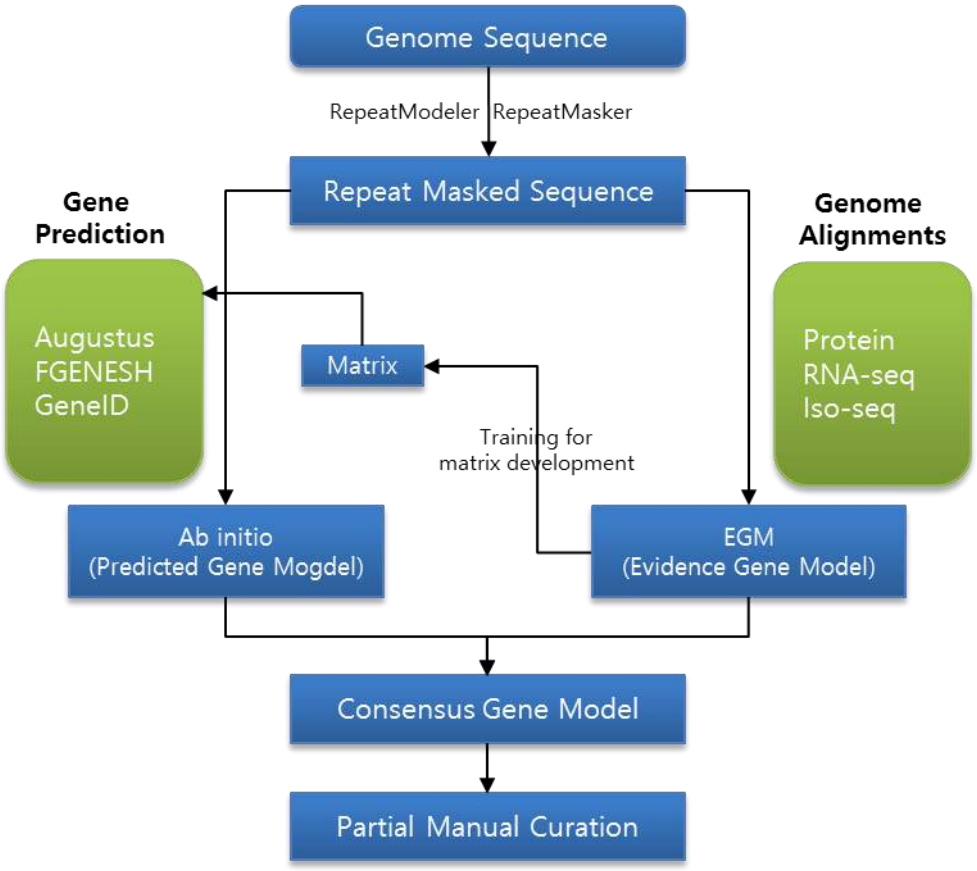



**Figure 3 - 2.** Gene prediction flowchart used for *Haliotis discus hannai* genome.

**Table 3 - 9.** Summary statistics of transcriptome read mapping using Tophat2.

| Categories | Tissue Samples | | | | | |
|---|---|---|---|---|---|---|
| | **Blood** | **Digestive duct** | **Gill** | **Hepatopancreas** | **Mantle** | **Ovary** |
| Total number of input pairs | 25,699,638 (100%) | 26,936,275 (100%) | 31,719,903 (100%) | 27,769,592 (100%) | 31,328,424 (100%) | 28,756,303 (100%) |
| Concordant pair mapping rate | 84.0% | 76.3% | 75.1% | 74.7% | 76.4% | 80.3% |
| Multiple alignments | 14.5% | 14.7% | 14.1% | 15.5% | 14.3% | 16.8% |
| Discordant alignment | 0.9% | 2.0% | 1.7% | 2.0% | 1.5% | 1.5% |
| Overall read mapping rate | 89.2% | 84.6% | 83.0% | 82.9% | 83.7% | 87.0% |

**Table 3 - 10.** Summary statistics RNA-seq based gene model using cufflinks.

| Element | Total count | Total length, bp | Mean length, bp |
|---|---|---|---|
| Transcript | 103,930 | 826,677,815 | 7,954.2 |
| Exon | 339,953 | 139,574,152 | 410.6 |

Gene prediction data from each method were combined using EVM (Evidence Modeler)(Haas et al. 2008) to build consensus gene set for abalone genome. All of gene models were converted to EVM compatible GFF3 format and merged to consensus gene set. After consensus gene annotation generated from EVM, manual curation was conducted for abandon genes from EVM to build final consensus gene set of *H.discus hannai*. Summary statistics for consensus gene set is shown in the Table 3-14 and Figure 3-3.

**C4. Functional Assignment**

All of the above gene evidences were combined using EVM(Haas et al. 2008), to build consensus gene set for abalone genome. To identify the function of predicted genes, I aligned protein sequences from consensus gene set to protein databases including SWISS-PROT(Consortium 2011), NCBI NR(Coordinators 2013) using Blastp (E-value < 1E-7, >50% sequence similarity, and >50% alignment coverage). Table 3-15 shows summary statistics for Blast2GO Result. Gene Ontology (GO) terms of each genes were obtained using the Blast2GO(Conesa et al. 2005) pipeline (Figure 3-4).

**Table 3 - 11.** Summary statistics of Iso-seq transcriptome data alignment using PASA pipeline.

| | **Input_seq** | **Mapped_seq** | **Mapping_rate** |
|---|---|---|---|
| *Haliotis discus hannai* | 40,993 | 40,627 | 99.11 |

**Table 3 - 12.** Summary statistics of protein alignment using tBlastn for the protein based evidence gene structure.

| Species | Type | Element | Total count | Count/ Gene | Total length, bp | Mean length, Bp | Genome Coverage % |
|---|---|---|---|---|---|---|---|
| *Homo sapiens* | Protein (69,002) | Transcript | 18,792 | | 109,068,639 | 5,803.99 | 5.80 |
| | | Exon | 77,320 | 4.11 | 12,667,395 | 163.83 | 0.67 |
| *Danio rerio* | Protein (42,474) | Transcript | 11,605 | | 68,796,463 | 5,928.17 | 3.66 |
| | | Exon | 47,300 | 4.08 | 7,978,167 | 168.67 | 0.42 |
| *Oncorhynchus mykiss* | Protein (53,876) | Transcript | 15,901 | | 55,043,032 | 3,461.61 | 2.93 |
| | | Exon | 46,040 | 2.90 | 7,567,059 | 164.36 | 0.40 |
| *Lottia gigantea* | Protein (23,851) | Transcript | 29,345 | | 177,851,531 | 6,060.71 | 9.47 |
| | | Exon | 118,165 | 4.03 | 20,583,999 | 174.20 | 1.10 |
| Croatea gigas | Protein (28,027) | Transcript | 32,978 | | 231,175,282 | 7,009.98 | 12.30 |
| | | Exon | 140,784 | 4.27 | 23,649,828 | 167.99 | 1.26 |
| *Aplysia californica* | Protein (29,096) | Transcript | 10,570 | | 67,396,621 | 6,376.22 | 3.59 |
| | | Exon | 45,737 | 4.33 | 7,797,503 | 170.49 | 0.42 |
| *Strongylocentrotus purpuratus* | Protein (38,730) | Transcript | 9,116 | | 46,270,640 | 5,075.76 | 2.46 |
| | | Exon | 34,572 | 3.79 | 5,627,082 | 162.76 | 0.30 |
| *Branchiostoma floridae* | Protein (58,493) | Transcript | 27,438 | | 125,307,206 | 4,566.92 | 6.67 |
| | | Exon | 92,426 | 3.37 | 15,483,164 | 167.52 | 0.82 |

**Table 3 - 13.** Summary statistics for ab initio gene prediction result using various programs and parameters

| Program | Matrix | Element | Total count | Count/Gene | Total length, bp | Mean length, bp | Genome Coverage % |
|---|---|---|---|---|---|---|---|
| Augustus | Custom parameter (RNAseq) | Gene | 88,825 | 3.92 | 367,066,732 | 4,132.47 | 19.54 |
| | | CDS | 348,528 | | 76,388,076 | 219.17 | 4.07 |
| | Custom parameter (*H.discus hannai* IsoSeq) | Gene | 90,396 | 4.11 | 395,511,710 | 4,375.32 | 21.05 |
| | | CDS | 371,487 | | 78,508,401 | 211.34 | 4.18 |
| | Custom parameter (H.discus discus IsoSeq) | Gene | 84,322 | 3.97 | 346,455,180 | 4,108.72 | 18.44 |
| | | CDS | 335,103 | | 72,527,841 | 216.43 | 3.86 |
| | Custom parameter (BUSCO) | Gene | 111,058 | 4.24 | 626,749,935 | 5,643.45 | 33.36 |
| | | CDS | 470,839 | | 84,333,972 | 179.11 | 4.49 |
| | Custom parameter (CEGAM) | Gene | 76,504 | 4.95 | 393,121,657 | 5,138.58 | 20.92 |
| | | CDS | 378,485 | | 63,424,677 | 167.58 | 3.38 |
| | Custom parameter (Protein) | Gene | 22,420 | 3.43 | 184,289,721 | 8,219.88 | 9.81 |
| | | CDS | 76,848 | | 20,291,739 | 264.05 | 1.08 |
| Fgenesh | Custom parameter | Gene | 184,051 | 3.46 | 1,366,924,540 | 7,426.88 | 72.75 |
| | | CDS | 636,568 | | 98,055,591 | 154.04 | 5.22 |
| Geneid | *Ciona intestinalis* | Gene | 789,540 | 1.41 | 436,990,370 | 553.47 | 23.26 |
| | | CDS | 1,112,959 | | 140,976,492 | 126.67 | 7.50 |

**Table 3 - 14.** Summary statistics for consensus gene set of *Haliotis discus hannai* genome.

| Element | No. elements | Exon/transcript | Avg. length | Total length | Genome coverage |
|---|---|---|---|---|---|
| Gene | 29,449 | - | 2,705 | 79,661,536 | 4.2 % |
| Exon | 74,745 | 2.54 | 280 | 20,985,298 | 1.1 % |
| Intron | 45,296 | 1.54 | 1,295 | 58,676,238 | 3.1 % |

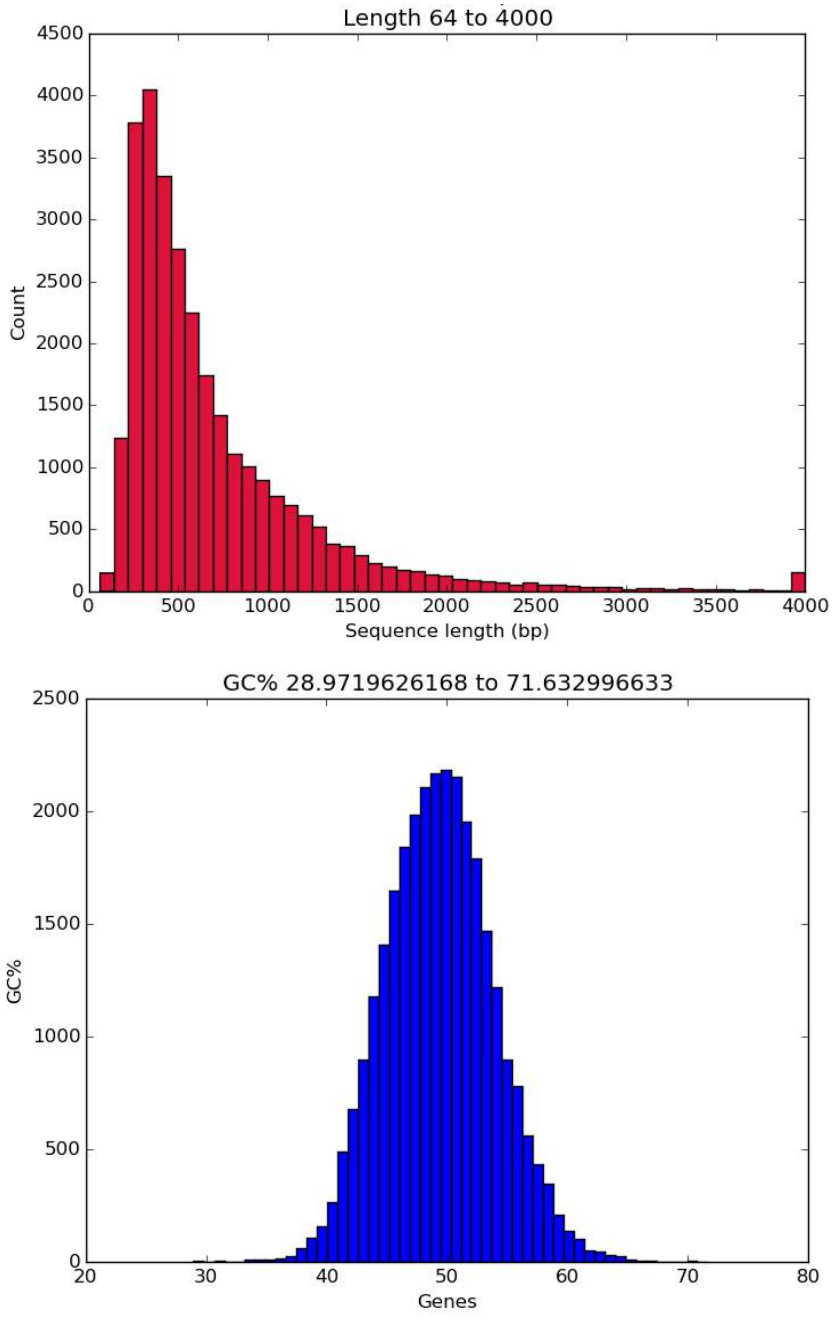



**Figure 3 - 3.** Histogram of (a)gene length and (b)GC contents for the consensus gene set of *Haliotis discus hannai* genome.

**Table 3 - 15.** Summary statistics for Blast2GO result.

| **Total** | **29,449** |
|---|---|
| No Hits | 7,916 (26.9%) |
| Blast hits | 21,533 (73.1%) |
| Gene Ontology Pathway | 6,880 |
| KEGG Pathway | 1,622 |

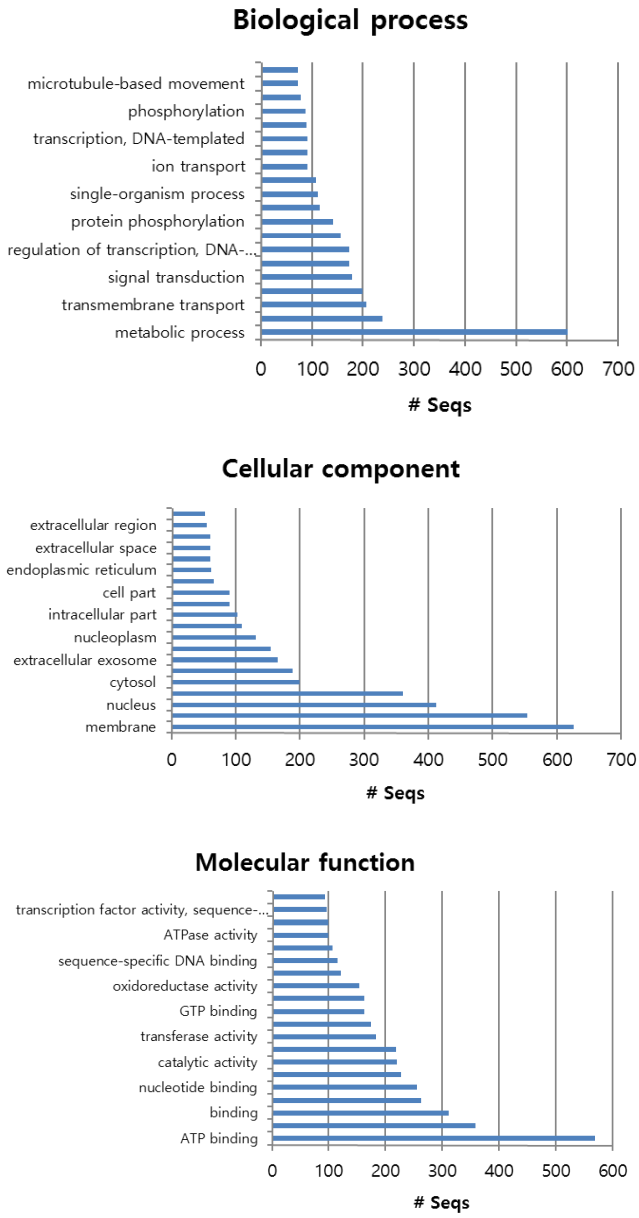



**Figure 3 - 4.** Number of genes distribution for top 10 related gene ontology (GO) term identified using Blast2GO.

## D. COMPARATIVE GENOME ANALYSIS

### D1. Orthologous gene set and Gene Family Group construction

First, I used mestortho(Kim et al. 2008) method for constructing 1:1 Orthologous gene set for 11 species. CDS and protein sequences of 11 species (*Homo sapiens, Mus musculus, Gallus gallus, Xenopus tropicalis, Danio rerio, Drosphila melanogaster, Caenorhabditis elegans, Capitella teleta, Crassostrea gigas, Lottia gigantean* and *Haliotis discus hannai*) were downloaded from Ensembl database. I conducted all-to-all pairwise alignment using AB-BLAST with E-value < 1E-7. A connection edge between two genes was assigned if query coverage and subject coverage were between 80% and 120%. BSR(Blast Score Ratio) was calculated. When 2 proteins P1 and P2 are compared, BSR=scoreP1P2/max (self-scoreP1 or self-scoreP2). Pairs with BSR > 1/3 were retained for clustering. E-values of remaining pairs were converted to values used in Ensembl ortholoy detection. (if E-value > 1e-199, int(round(-math.log10(E-value)/2 and otherwise, 100). Based on these values, a gene cluster was built by hcluster_sg used in TreeFam(Ruan et al. 2008). Using minimum evolution algorithm of Mestortho, paralogous genes were removed from homologous genes and 53 genes, in total, were determined as 1:1 orthologous gene.

Second, I used OrthoMCL(Li et al. 2003) for gene family group analysis. CDS and protein sequences of 3 close neighbors (*Capitella teleta, Crassostrea gigas* and *Lottia gigantea*) of *H.discus hannai*, and available marine genomes in Ensembl database, were downloaded. Following the guide instruction of OrthoMCL, a total of 19,306 gene family groups was determined for downstream analysis.

**D2. Phylogenetic Tree**

To build the species phylogenetic tree based on the 1:1 orthologous genes of 11 species, multiple sequence alignment for each gene was conducted using PRANK(Löytynoja and Goldman 2005). Gblocks(Castresana 2000) was used to obtain the conserved blocks from the multiple sequence alignment result. Conserved block sequences were sequentially concatenated to one consensus sequence for each species. Next, MEGA6(Tamura et al. 2013) was used for constructing Neighbor-Joining Tree of 11 species with bootstrap 1000, maximum composite likelihood, transitions + transversions, and gamma distributed option. Species tree is shown in Figure 3-5.

**D3. Gene Family Extension and Contraction Analysis**

Number of genes in each gene family group were counted and converted to input data for CAFE(De Bie et al. 2006) analysis. Divergence time was estimated using average divergence time in Timetree(Hedges et al. 2006) based on the previous studies(Peterson et al. 2008, Vinther et al. 2012). Estimated divergence time is 494, 532 and 574 million years ago between *H.discus hannai* and *L. gigantea*, *C. gigas, C. teleta*, respectively. CAFE analysis was conducted using one parameter model and pathway analysis for expanded gene families of *H.discus hannai* was conducted using DAVID(Huang et al. 2009). Enriched Gene Ontology term is summarized in Table 3-16.

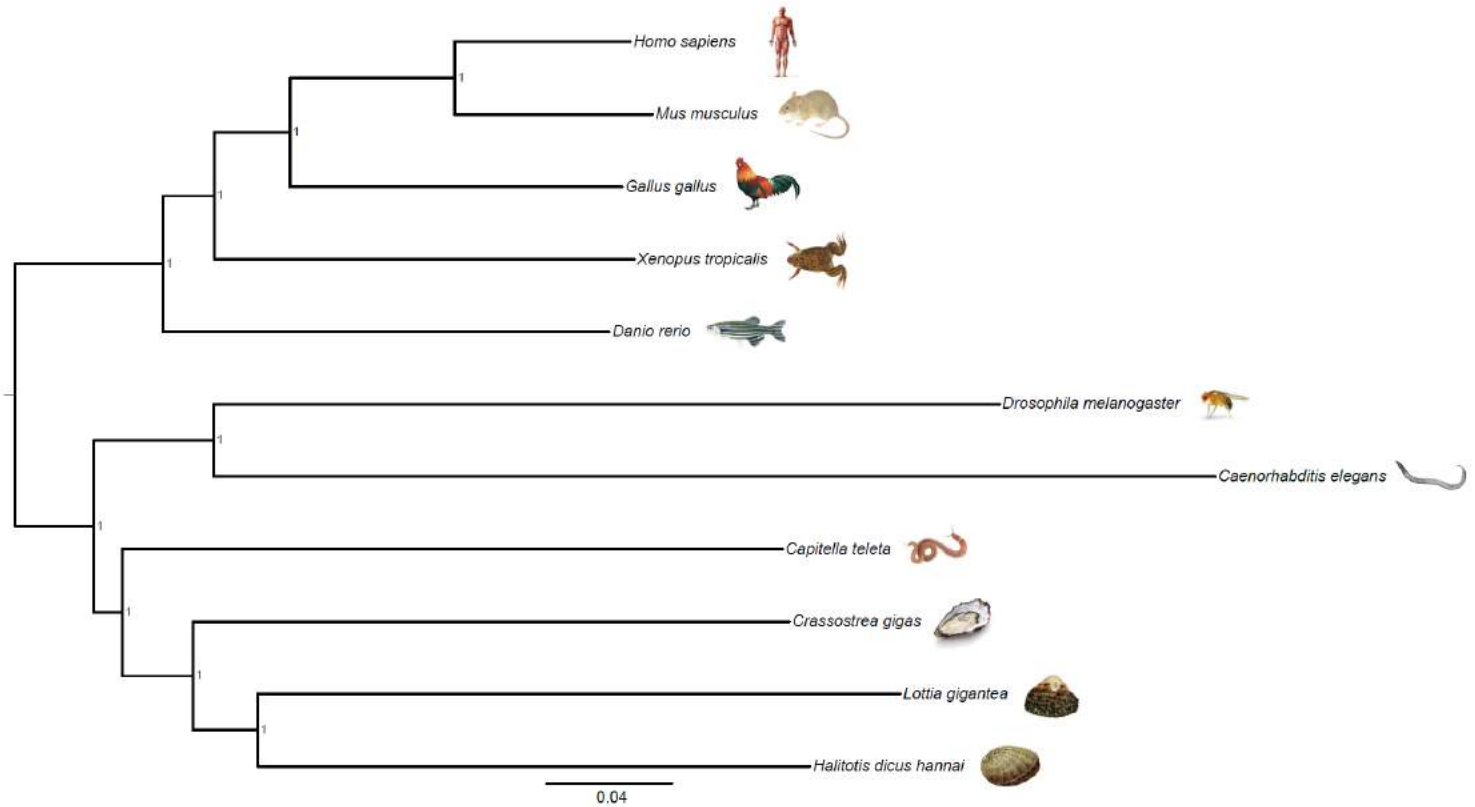



**Figure 3 - 5.** Neighbor Joining Tree using 53 orthologous genes for 11 species. Node values indicate bootstrap value.

**Table 3 - 16.** Enriched gene ontology terms related with expanded genes families in *Haliotis discus hannai* genome using DAVID pathway analysis. (p-value < 0.05, BP – Biological Process, CC – Cellular Component, MF – Molecular Function)

| Term ID | Description | Type | P-value |
|---|---|---|---|
| GO:0007155 | cell adhesion | BP | 1.26E-05 |
| GO:0022610 | biological adhesion | BP | 1.28E-05 |
| GO:0006814 | sodium ion transport | BP | 0.001138 |
| GO:0008202 | steroid metabolic process | BP | 0.00348 |
| GO:0008206 | bile acid metabolic process | BP | 0.009923 |
| GO:0007016 | cytoskeletal anchoring at plasma membrane | BP | 0.014554 |
| GO:0006897 | endocytosis | BP | 0.017079 |
| GO:0010324 | membrane invagination | BP | 0.017079 |
| GO:0045765 | regulation of angiogenesis | BP | 0.017307 |
| GO:0055085 | transmembrane transport | BP | 0.018863 |
| GO:0016127 | sterol catabolic process | BP | 0.022664 |
| GO:0006707 | cholesterol catabolic process | BP | 0.022664 |
| GO:0032102 | negative regulation of response to external stimulus | BP | 0.031389 |
| GO:0050728 | negative regulation of inflammatory response | BP | 0.037314 |
| GO:0015672 | monovalent inorganic cation transport | BP | 0.043027 |
| GO:0051276 | chromosome organization | BP | 0.044201 |
| GO:0007059 | chromosome segregation | BP | 0.044769 |
| GO:0009416 | response to light stimulus | BP | 0.048242 |
| GO:0006261 | DNA-dependent DNA replication | BP | 0.049945 |
| GO:0007588 | excretion | BP | 0.049945 |
| GO:0031430 | M band | CC | 0.001287 |
| GO:0031672 | A band | CC | 0.007501 |
| GO:0031012 | extracellular matrix | CC | 0.008516 |
| GO:0005578 | proteinaceous extracellular matrix | CC | 0.009291 |
| GO:0043292 | contractile fiber | CC | 0.013336 |
| GO:0030017 | sarcomere | CC | 0.013595 |
| GO:0005694 | chromosome | CC | 0.016668 |
| GO:0030016 | myofibril | CC | 0.025199 |
| GO:0044449 | contractile fiber part | CC | 0.027446 |
| GO:0044420 | extracellular matrix part | CC | 0.032341 |

| | | | |
|---|---|---|---|
| GO:0043167 | ion binding | MF | 1.20E-04 |
| GO:0043169 | cation binding | MF | 1.37E-04 |
| GO:0046872 | metal ion binding | MF | 3.15E-04 |
| GO:0005509 | calcium ion binding | MF | 8.11E-04 |
| GO:0015293 | symporter activity | MF | 8.60E-04 |
| GO:0031402 | sodium ion binding | MF | 9.07E-04 |
| GO:0030169 | low-density lipoprotein binding | MF | 0.001861 |
| GO:0008034 | lipoprotein binding | MF | 0.001932 |
| GO:0015294 | solute:cation symporter activity | MF | 0.003189 |
| GO:0008307 | structural constituent of muscle | MF | 0.004369 |
| GO:0005085 | guanyl-nucleotide exchange factor activity | MF | 0.006181 |
| GO:0008146 | sulfotransferase activity | MF | 0.008459 |
| GO:0030246 | carbohydrate binding | MF | 0.011068 |
| GO:0051393 | alpha-actinin binding | MF | 0.012867 |
| GO:0031420 | alkali metal ion binding | MF | 0.015712 |
| GO:0016782 | transferase activity, transferring sulfur-containing groups | MF | 0.016851 |
| GO:0005201 | extracellular matrix structural constituent | MF | 0.023948 |
| GO:0042805 | actinin binding | MF | 0.026195 |
| GO:0005539 | glycosaminoglycan binding | MF | 0.029117 |
| GO:0004197 | cysteine-type endopeptidase activity | MF | 0.038537 |
| GO:0005198 | structural molecule activity | MF | 0.040184 |
| GO:0015370 | solute:sodium symporter activity | MF | 0.045155 |
| GO:0001871 | pattern binding | MF | 0.046922 |
| GO:0030247 | polysaccharide binding | MF | 0.046922 |

## E. COMPARATIVE GENOME ANALYSIS WITHIN PACIFIC ABALONE SPECIES

### E1. Sample Collection for Comparative analysis

The details on samples collections are as follows:

1. *Haliotis discus hannai*, from Professor Yamakawa of TUMST in March, 2015.
2. *Haliotis gigantea,* from Professor Yamakawa of TUMST in March, 2015.
3. *Haliotis makadaka,* from Professor Lim of Mokpo University in April, 2015.

### E2. Library Construction and Data Generation

For genomic DNA extraction, hemocytes were harvested from fresh hemolymph by centrifugation at 3000 × rpm for 5 min at 4℃. Genomic DNA was extracted using a DNeasy Animal Mini Kit (Qiagen, Hilden, Germany). All sequencing processes were performed following manufacturer instructions same as draft genome assembly. Generated whole genome sequencing data of 3 species summarized in Table 3-17.

**Table 3 - 17.** Summary statistics of generated whole genome sequencing data for 3 pacific abalone species.

| Library Name | Library Type | Insert Size | Platform | Read Length | No. Read | Total bp |
|---|---|---|---|---|---|---|
| *Haliotis discus discus* | Paired-end | 350 | Hiseq2000 | 101 | 439,252,210 | 44,364,473,210 |
| *Haliotis madaka* | Paired-end | 350 | Hiseq2000 | 101 | 442,530,884 | 44,695,619,284 |
| *Haliotis gigantea* | Paired-end | 350 | Hiseq2000 | 101 | 467,486,420 | 47,216,128,420 |
| **Total** | | | | | | 136,276,220,914 |

**E3. Variant Calling and Annotation**

The quality of the whole genome sequencing reads from all samples was checked using FastQC(Andrews 2010) and raw data quality control process was conducted using Trimmomatic(Bolger et al. 2014) with the parameter (ILLUMINACIP:TruSeq3-PE-2.fa:2:30:10, MINLEN:75). Paired-end reads of each sample were mapped to the draft genome of H.discus hannai using Bowtie2 (Langmead and Salzberg 2012) with the default options (Table 3-18). Using the "REMOVE_DUPLICATES=true" option in "MarkDuplicates" command-line tool of Picard (http://picard.sourceforge.net, v.1.109), potential PCR duplicates were removed. I then used SAMtools (Li et al. 2009) to create index files for reference and bam files. Genome Analysis Toolkit (GATK) (McKenna et al. 2010) was used for downstream processing and variant calling. Local realignment was conducted using GATK to correct misalignments due to the presence of indels. The "Haplotypecaller" of GATK was used for calling candidate single nucleotide variants (SNVs) and InDels. To avoid possible false positive variants, argument "VariantFiltration" of the same software was adopted with the following options: 1) SNVs with a phred-scaled quality score of less than 30 were filtered; 2) SNVs with QD(unfiltered depth of non-reference samples, low scores indicate false positives and artifacts) <5 were filtered; 3) SNVs with FS (phred-scaled P value using

Fisher's exact test) >200 were filtered as FS represents variation on either the forward or the reverse strand, which are indicative of false-positive calls; 4) SNVs with MQ0(the number of reads which have mapping quality zero) >4 or MQ0/DP(proportion of mapping quality zero reads over total depth) >0.1 were filtered to remove uncertain calls; 5) more than 3 SNVs within 10bp window were filtered. Likewise, I also excluded variants that were predicted to overlap with InDel mutation, or overlap with zero coverage regions in more than one sample. Variant annotation was conducted through SnpEff(Cingolani et al. 2012), based on the custom library built from consensus gene set. Number of identified variants and annotation information is shown in Figure 3-6, 3-7.

**E4. MT genome sequencing reconstruction**

Whole mitochondria genome sequences, of 4 pacific abalone species, were reconstructed using variants from each sample based on the sequence of the complete mitochondria genome sequence of *H.discus hannai* (gi|564733422|gb|KF724723.1|). Paired-end reads of each sample were mapped to the complete MT genome sequence using Bowtie2 (Langmead and Salzberg 2012) and same procedure was employed in whole genome variant calling analysis. Using the complete MT genome sequence of *H.discus hannai*

**Table 3 - 18.** Summary of whole genome read mapping using paired-end read with Bowtie2.

| Categories | *Haliotis discus hannai* | *Haliotis discus discus* | *Haliotis madaka* | *Haliotis gigantea* |
|---|---|---|---|---|
| Total number of reads pairs | 423,350,776 (100%) | 207,517,395 (100.00%) | 206,602,362 (100.00%) | 214,065,416 (100.00%) |
| Concordantly 1 time | 177,892,969 | 21,502,612 | 17,365,776 | 14,585,851 |
| Concordantly > 1 time | 199,497,044 | 43,177,517 | 39,906,281 | 36,499,008 |
| Discordantly 1 time | 3,307,932 | 38,850,688 | 41,708,829 | 39,162,423 |
| 1 time in mixed mode (singleton) | 14,886,969 | 37,731,746 | 38,944,492 | 45,711,482 |
| > 1 time in mixed mode (singleton) | 27,135,765 | 132,222,353 | 140,777,533 | 140,330,677 |
| Overall mapping rate including singletons | 94.89% | 90.84% | 91.40% | 85.61% |

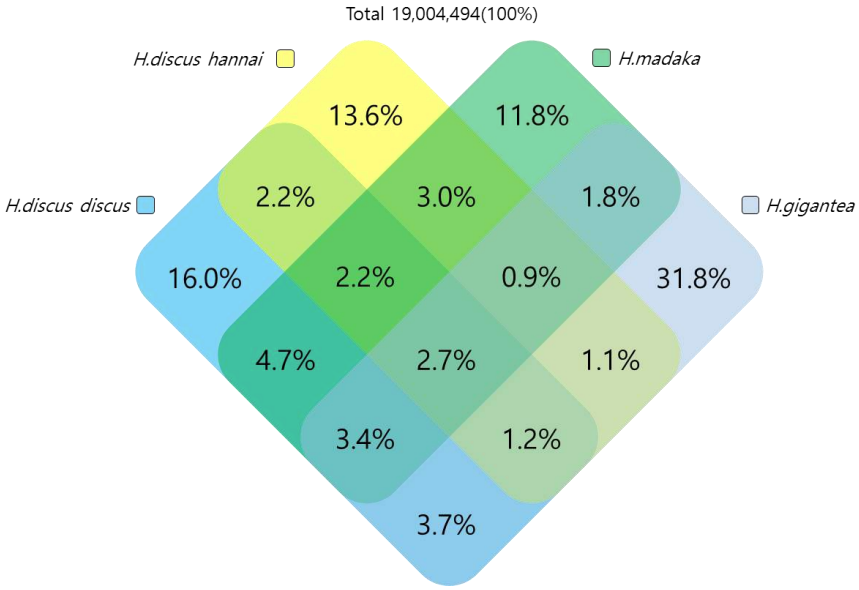



**Figure 3 - 6.** Number of variant from whole genome re-sequencing using 3 different Haliotis species with *H.discus hannai*.

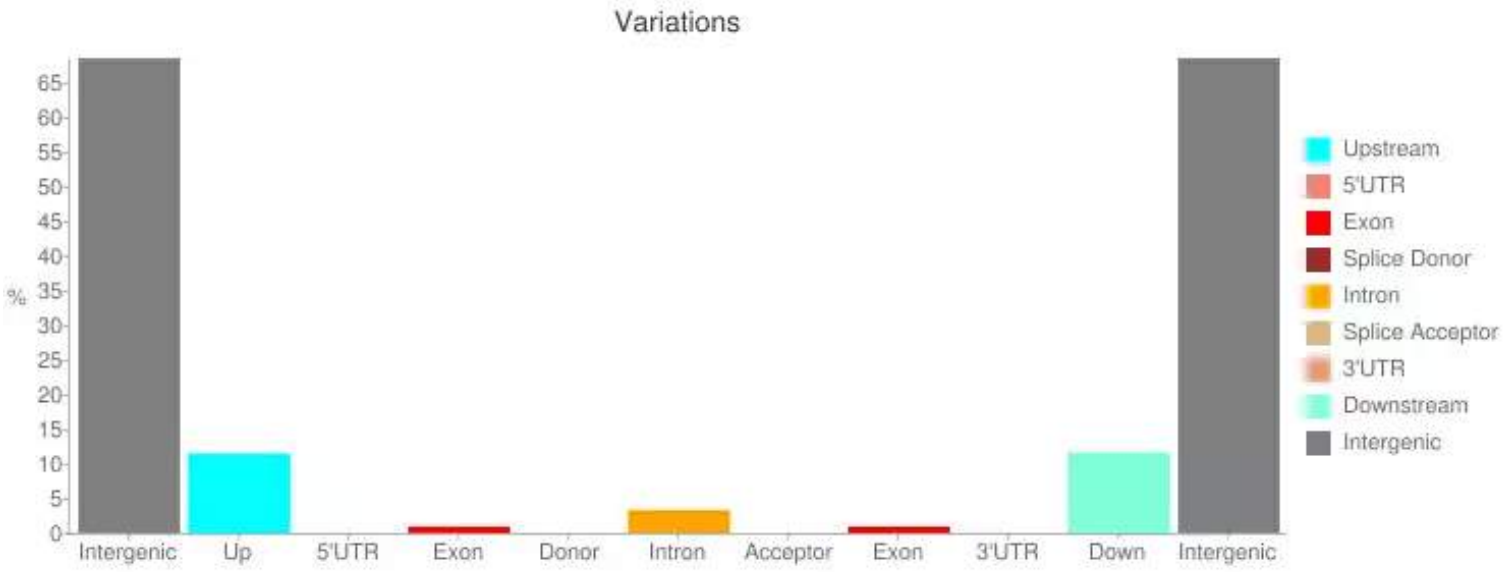



**Figure 3 - 7.** Regional distribution of whole genome variants of 4 different Haliotis species using SnpEff.

as a backbone sequence, reliable SNVs and InDels were converted for each species.

**E5. Divergence Time estimation of Pacific Abalone Species**

3 complete MT genome sequences of other Haliotidae species (*H. laevigata* : gi|669026534|ref|NC_024562.1|, *H. rubra* : gi|49146557|ref|NC_005940.1| and *H. tuberculata tuberculata : gi|220898237|gb|FJ599667.1|*) were download from NCBI and used for outgroup of 4 pacific abalone species. Multiple sequence alignment was conducted using PRANK(Löytynoja and Goldman 2005). To obtain the conserved blocks from the multiple sequence alignment result, Gblocks(Castresana 2000) was used for data preparation. Modeltest(Posada and Crandall 1998) was carried out to determine a proper evolution model. The results of the likelihood scores for each of the 56 models using PAUP(Swofford 2003) were used as input data for Modeltest. Best evolution model for pacific abalone MT genome was selected based on the results of Modeltest using AIC (Akaike Information Criterion). To estimate the divergence time of 4 pacific abalone species, Bayesian phylogenetic analysis was conducted(Bouckaert et al. 2014). GTR substitution model was selected based on the Modeltest result, and clock model was set to Relaxed Clock Log Normal. Calibrated Yule Model was employed and the prior and

sigma value was set to 100.5 and 5.0 respectively. MCMC chain length was set to 100 million times.

**E6. Demographic History of Pacific Abalone.**

I estimated the demographic history of 4 pacific abalone species using the pairwise sequentially Markovian coalescent method(Li and Durbin 2011). Population size history from a diploid sequence can be inferred using PSMC. Consensus sequence data was generated from alignment read data using samtools(Li et al. 2009) with parameters -d 10, -d 100, and bootstrap sampling was executed 100 times for each sample. For the resulting plots, generation time was set to 3 years based on the previous study(Shepherd et al. 1992). Average mutation rate per nucleotide of abalone species was not yet known, I used default average nucleotide mutation rate.

## F. COMPARATIVE POPULATION ANALYSIS

### F1. Sample Collection

Sample preparation, for GBS analysis on abalone population study, was done according to the following descriptions:

1. Goseong, Korea (24 *hannai* samples collected in July, 2013; shell length 9¡¾2 cm).

2. Yeosu, Gangwon-do, Korea (30 *hanai* samples collected in August, 2014; shell length 10¡¾1 cm)

3. Taean, Korea (33 *hannai* samples collected in August, 2014; )

4. Mie, Japan (5 *discus* samples provided by Prof. Yamakawa of TUMST in March, 2015; shell length 11¡¾1 cm)

5. Nagasaki, Japan (1 *discus* sample provided by Prof. Yamakawa of TUMST in March, 2015; shell length 12 cm)

6. Nagasaki, Japan (1 *gigantea* sample provided by Prof. Yamakawa of TUMST in March 2015; shell length 14 cm)

7. California, USA (11 *rufescens* samples provided by Prof. Lim of Mokpo Univ. in October, 2015)

**F2. Library Construction and Data Generation**

Total genomic DNA was extracted from muscle tissue using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instruction. The amount of DNA was quantified using the standard procedure of Quant-iT PicoGreen dsDNA Assay Kit (Molecular Probes, Eugene, OR, USA) with Synergy HTX Multi-Mode Reader (Biotek, Winooski, VT, USA) and normalized to 20 ng/ul. DNA (200ng) was digested with 8U of High-fidelity PstI (New England BioLabs, Ipswich, MA, USA) at 37 C for 2 hours and heated to 65 C for 20 minutes to inactivate the enzyme.

DNA libraries for genotyping-by-sequencing (GBS) were constructed according to the protocols as described previously (Elshire et al. 2011, De Donato et al. 2013) with minor modifications. The restriction digestion of DNA with PstI was followed by ligation with adapters. The sets of 105 ligations were purified using QIAquick PCR Purification Kit (Qiagen). The pooled ligations (5ul) were amplified in 50ul reaction by multiplexing PCR using AccuPower Pfu PCR Premix (Bioneer, Daejeon, South Korea) and 25 pmol of each primer below:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA
CGCTCTTCCGATCT - 3'

and

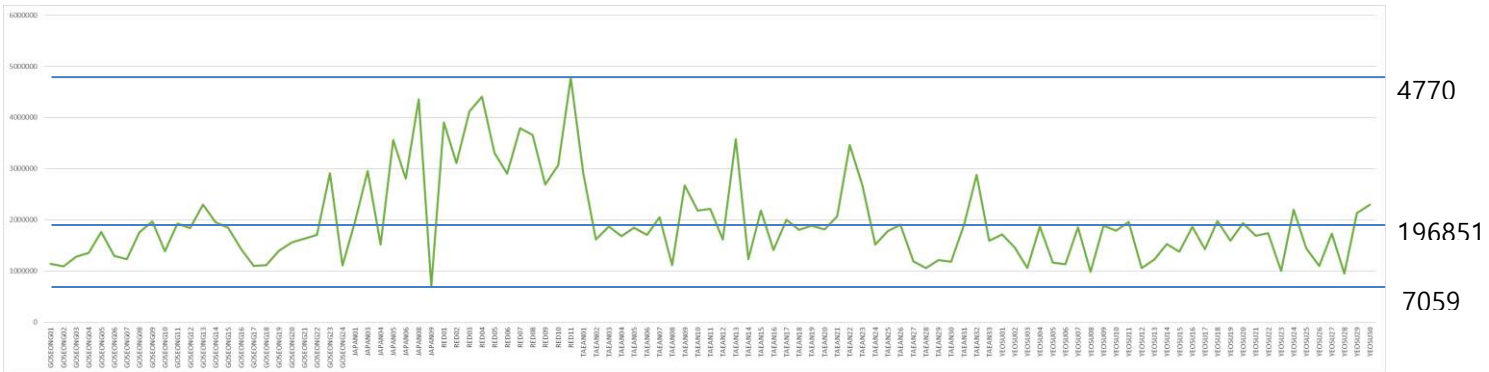
5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTG AACCGCTCTTCCGATCT - 3’.

PCR cycles consisted of 98 C for 5 min followed by 18 cycles of 98 C for 10 s, 65 C for 5 s, and 72 C for 5 s, with a final extension step at 72 C for 5 min. The PCR products were also purified using QIAquick PCR Purification Kit (Qiagen) and then evaluated the distribution of fragment sizes with BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). The GBS libraries were sequenced in the Illumina NextSeq500 (Illumina, San Diego, CA, USA) with the length of 150 bp single-end reads following the manufacturer instruction. Read count distribution of 105 GBS libraries is shown in Figure 3-8.

**F3. Variant Calling**

Before conducting variant calling analysis, I conducted de-multiplexing of GBS data using GBXtoolskit(Herten et al. 2015). Reads of each sample were mapped to the draft genome of *H.discus hannai* using Bowtie2 (Langmead and Salzberg 2012) with the default options. I then used SAMtools (Li et al. 2009) to create index files for reference and bam files. Genome Analysis

Toolkit (GATK) (McKenna et al. 2010) was used for downstream processing and variant calling. Local realignment was conducted using GATK to correct misalignments due to the presence of InDels. The “Haplotypecaller” of GATK was used for calling candidate single nucleotide variants (SNVs) and InDels. To avoid possible false positive variants, argument “VariantFiltration” of the same software was adopted with the following options: 1) SNVs with a phred-scaled quality score of less than 30 were filtered; 2) SNVs with QD (unfiltered depth of non-reference samples; low scores are indicative of false positives and artifacts) <5 were filtered; 3) SNVs with FS (phred-scaled P value using Fisher’s exact test) >200 were filtered, as FS represents variation on either the forward or the reverse strand, which are indicative of false-positive calls; 4) SNVs with MQ0 (the number of reads which have mapping quality zero) >4 or MQ0/DP (proportion of mapping quality zero reads over total depth) >0.1 were filtered to remove uncertain calls; 5) more than 3 SNVs within 10bp window were filtered. Likewise, I also excluded variants that were predicted to overlap with InDel mutation, or overlap with zero coverage regions in more than one sample.




Species: *Haliotis discus hannai, Haliotis discus discus, Haliotis gigantean, Haliotis rufenscens*

**Figure 3 - 8.** De-multiplexed read count distribution of 105 GBS libraries of Haliotidae population.

**F4. Principal Component Analysis and Population Structure Analysis**

Principal component analysis, based on the SNV information of 105 samples, was conducted using SNPRelate package(Zheng et al. 2012). For population admixture analysis, VCF file converted to plink file format using PGSpider(Lischer and Excoffier 2012) and 61,795 SNVs were randomly sampled using plink(Purcell et al. 2007) with --thin option. Randomly sampled variants were converted to STRUCTURE input again and population admixture was estimated using STRUCTURE(Hubisz et al. 2009) with 5,000 burn-in and 50,000 MCMC iteration.

## G. METAGENOME AND META-TRASCRIPTOME ANALYSIS

### G1. Sample Collection

The *H. discus hannai* (shell length 8 cm) samples, used in the metagenome analysis of abalone digestive system, were from purchased from Wan island, Korea in February, 2016. Only the alimentary canal, extracted through viscera dissection, was used in the analysis.

### G2. Library Construction and Data Generation

To amplify the variable V3 and V4 region of the 16s rRNA, I initially ran the 1st PCR accordingly: 1) 10 ng of DNA and 2x quick taq HS dye mix Z (Toyobo) were used to make 30ul reaction volume to perform PCR. 2) The PCR conditions followed a cycle of 95°C for 3 minutes; 25 cycles of 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds; 3) This was followed by the last extension of 72°C for 5 minutes and hold in 4°C. Also, the primers used in this analysis is as follows:

1st PCR Forward Primer =

5' - TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGG NGGCWGCAG - 3’

1st PCR Reverse Primer =

5' - GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHV GGGTATCTAATCC - 3’

The PCR products were then sequenced through Illumina Miseq 300bp paired-end library, using standard protocol, and the descriptions of the data is summarized in Table 3-19.

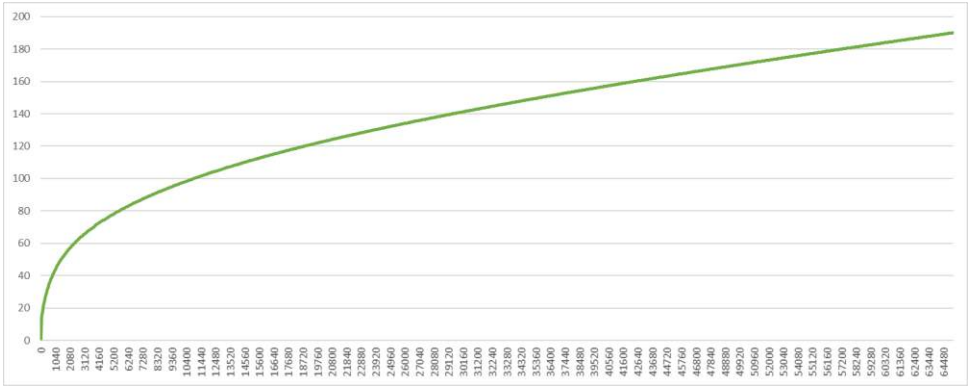
WTS library was constructed via illumina Truseq stranded total RNA sample preparation kit. The rRNAs were excluded from the total RNA through rRNA removal kit, followed by RNA purification with RNA Clean xp bead (Beckman Coulter). Then the remaining RNA is fragmented via EFP Mix (under PCR condition of 94°C for 6 minutes and 4°C hold). The fragmented RNAs are synthesized to 1st cDNA using Reverse transcriptase and random primers (under PCR condition of 25°C for 10 minutes, 42°C for 15 minutes, 70°C for 15 minutes, 4°C hold). For ds cDNA (2nd cDNA) synthesis, from resulting 1st cDNAs, second strand marking Master (SMM) was utilized (under PCR condition of 16°C for 1hour, 4°C hold). This library construction, through Illumina Nextseq500, followed the manufacturer’s protocol on data generation, and the details are provided in Table 3-19.

**G3. 16s Community Analysis**

Quality of generated reads were checked using FASTQC(Andrews 2010) and adapter sequences were removed using Trimmomatic(Bolger et al. 2014). Paried-end reads were merged using FLASH(Magoč and Salzberg 2011) to generate single 16s rRNA read. Finally, community analysis was conducted using MG-RAST(Glass et al. 2010) with Greengene database(DeSantis et al. 2006). Summary statistics for merged paired-end data is shown in Table 3-20. α-Diversity was 9.812 and rare fraction curve is shown in Figure 3-9. Figure 3-10 shows identified community information of abalone visceral extract in various level.

**G4. Meta-transcriptome Analysis**

Quality of generated reads were checked using FASTQC(Andrews 2010) and adapter sequences were removed using Trimmomatic(Bolger et al. 2014). I used the Trinity(Grabherr et al. 2011) *de novo* assembly tool to identify genes expressed in the visceral extract of abalone. Summary statistics for total transcriptome using Trinity is shown in Tabls S21. Function of identified transcripts were annotated using Swiss-Prot database using Blastx (e-value < 1e-5, >50% sequence similarity and >50% alignment coverage). Using paired-end read remapping, I filtered out the candidate miss assembled transcripts which showed low paired-end read mapping from meta-transcript assembly.




α-Diversity = 9.812 species

**Figure 3 - 9.** Rarefaction curve of annotated species richness in visceral extract of *Haliotis discus hannai*

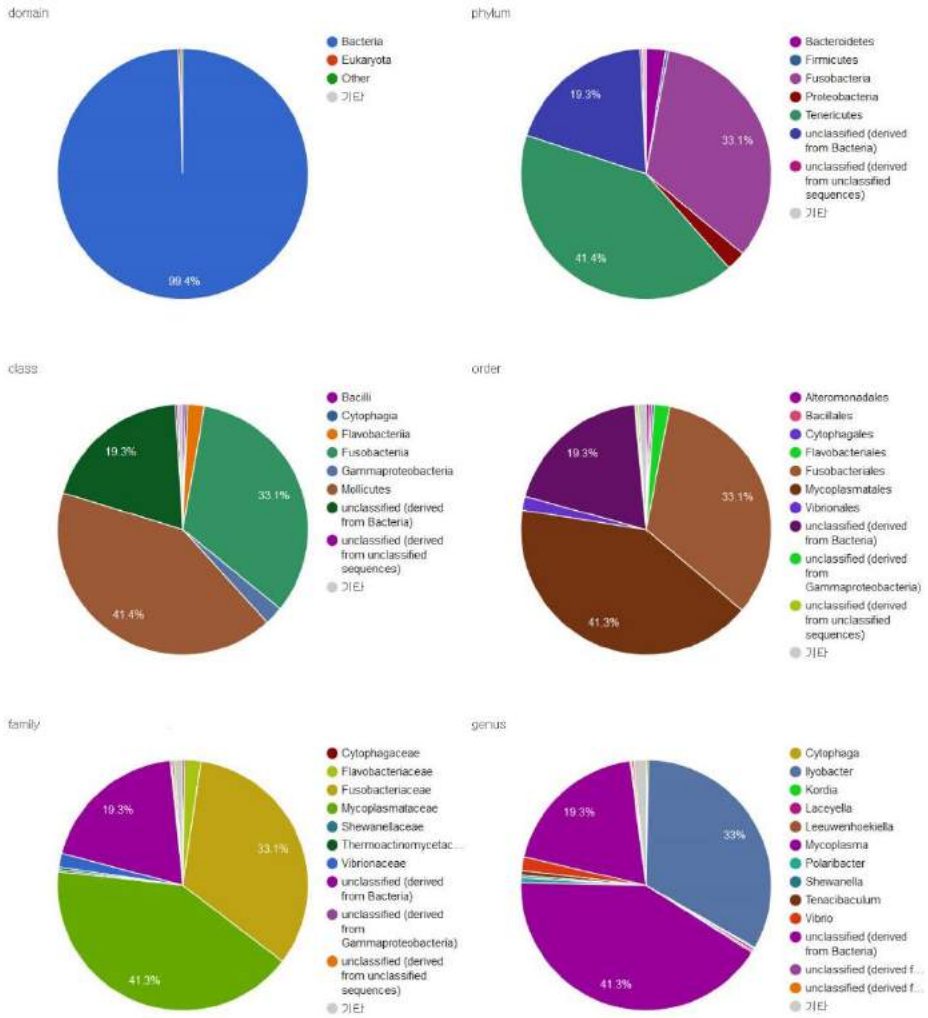



**Figure 3 - 10.** Pie chart for the distribution of taxanomic annotation using MG-RAST.

**Table 3 - 19.** Summary statistics of generated whole genome sequencing data for 3 pacific abalone species.

| Library Name | Library Type | Insert Size | Platform | Read Length | No. Read | Total bp |
|---|---|---|---|---|---|---|
| 16s amplicon | Paired-end | 550 | Miseq | 301 | 526,742 | 15,8549,342 |
| *Total Transcriptome* | Paired-end | 350 | Nextseq500 | 76 | 214,087,970 | 16,270,685,720 |

**Table 3 - 20.** Summary statistics for 16s rRNA library data used in community analysis of visceral extract of *Halitotis discus hannai*.

| **Input Sequence** | |
|---|---|
| bp Count | 29,383,116 bp |
| Sequences Count | 65,139 |
| Mean Sequence Length | 451 ± 14 bp |
| Mean GC percent | 49 ± 1 % |
| **Post Quality Control** | |
| bp Count | 29,172,480 bp |
| Sequences Count | 64,779 |
| Mean Sequence Length | 450 ± 11 bp |
| Mean GC percent | 49 ± 1 % |
| Predicted rRNA Features | 424 |
| Identified rRNA Features | 404 |

**Table 3 - 21.** Summary statistics for total transcriptome assembly of visceral extract using Trinity.

| | |
|---|---|
| Total trinity 'genes': | 343,970 |
| Total trinity transcripts: | 419,032 |
| Percent GC | 41.19 |
| Stats based on ALL transcript contigs | |
| Contig N50 | 1,156 |
| Median contig length | 364 |
| Average contig | 703.14 |
| Total assembled bases | 294,637,012 |
| Stats based on ONLY LONGEST ISOFORM per GENE | |
| Contig N50 | 714 |
| Median contig length | 339 |
| Average contig | 566.1 |
| Total assembled bases | 194,720,276 |

## 3.3 Results and Discussion

### Sequencing and Assembly

I employed whole genome shotgun approach to reconstruct and explore genetic features of *H.discus hannai* genome with various sequencing platforms. For assembly of *H.discus hannai* genome, I generated a total of 580G base pairs and it was 322-fold coverage based on the 1.8Gb, estimated genome size of *H.discus hannai* using flow cytometry. Constructed libraries, sequencing platforms, and more detailed information for data generation are in the Supplementary Information A and Table S1. Before genome assembly, I identified the characters of *H.discus hannai* genome and estimated the genome size using k-mer distribution (Figure 3-1). 19-mer distribution using the paired-end library with 350bp insert size showed that second peak exists in multiplicity value 23, the half of main peak in multiplicity value 47. These values indicate that the genome of *H.discus hannai* had high heterozygous characters, and I adjusted the assembly parameters to reflect these characteristics. Estimated genome size of *H.discus hannai* genome was 1.6Gb based on the 19-mer distribution and is smaller than the previous estimate, 1.8Gb, from flow cytometry(Adachi and Okumura 2012). I conducted numerous assemblies with varying assemblers and parameters, then selected the final assembly based on the result of Core Eukaryotic Genes Mapping

Approach (CEGMA) evaluation. The final assembly comprised of 1.86 Gb with 80,032 scaffolds, which is bigger than the genome size estimated from flow cytometry. I assume this resulted from assembly redundancy caused by the high heterozygosity of *H.discus hannai* genome. GC contents level was 40.47%, and N50 length of assembled scaffolds was about 200kb. Summary statistics for the *H.discus hannai* draft genome is depicted in Table 3-2. To evaluate the quality of *H.discus hannai* draft genome, I employed paired-end reads remapping, CEGMA and BUSCO (Benchmarking Universal Single-Copy Orthologs). 94.89% of paired-end read with 350bp insert size were successfully mapped to the assembled genome (Table 3-3). Assembled genome contains 95 complete and 198 partial genes in CEGMA analysis and 609 complete and 130 fragmented genes in BUSCO analysis. The detailed information on evaluation results is summarized in Table 3-4 and Table 3-5.

**Gene annotation and Hox gene atomization**

A total of 29,449 genes were predicted in the *H.discus hannai* genome using *ab-initio* gene prediction: protein evidence-based search and RNA-seq from multiple organs. The consensus gene set of *H.discus hannai* contains 74,745 exons and 45,296 introns, and the average length of exon and intron was 280bp and 1,295bp, respectively. Predicted genes covered 4.2% of the draft genome

(Table 3-14), and gene length and GC contents distribution is shown in the Figure 3-3. Among 29,449 predicted genes, 21,533 (73.1%) genes entries were matched to a known database using Blast2GO. Also, 6,680 gene ontology terms and 1,622 KEGG pathways were identified based on functional assignment (Table 3-15), then the top 10 related gene ontology terms are shown in Figure 3-4.

Based on our gene annotation, Hox gene cluster, which is known to play a major role in the developmental stage, was identified in *H.discus hannai* genome (Figure 3-11). Interestingly, Hox genes of *H.discus hannai* was completely atomized except for two Hox genes in scaffold_00168, and this is concordant to octopus, a previously studied mollusk(Albertin et al. 2015). The Hox gene cluster is controlled through cis-regulation if its ancestral type is preserved, as in *L.gigantea(Simakov et al. 2013)*. However, when the cluster is split, as in *C.gigas(Zhang et al. 2012)*, the Hox gene cluster is Cis-trans regulated and has been reported to be trans-regulated when further split to an atomized form(Duboule 2007). A recent study also reported that Hox genes of *Gibbula varia*, one of marine sea snails included in Gastropoda, expressed non-collinearly during developmental stage(Fritsch et al. 2015). These results indicate that this type of non-collinear expression of Hox gene, by cluster atomization in the genome, can be

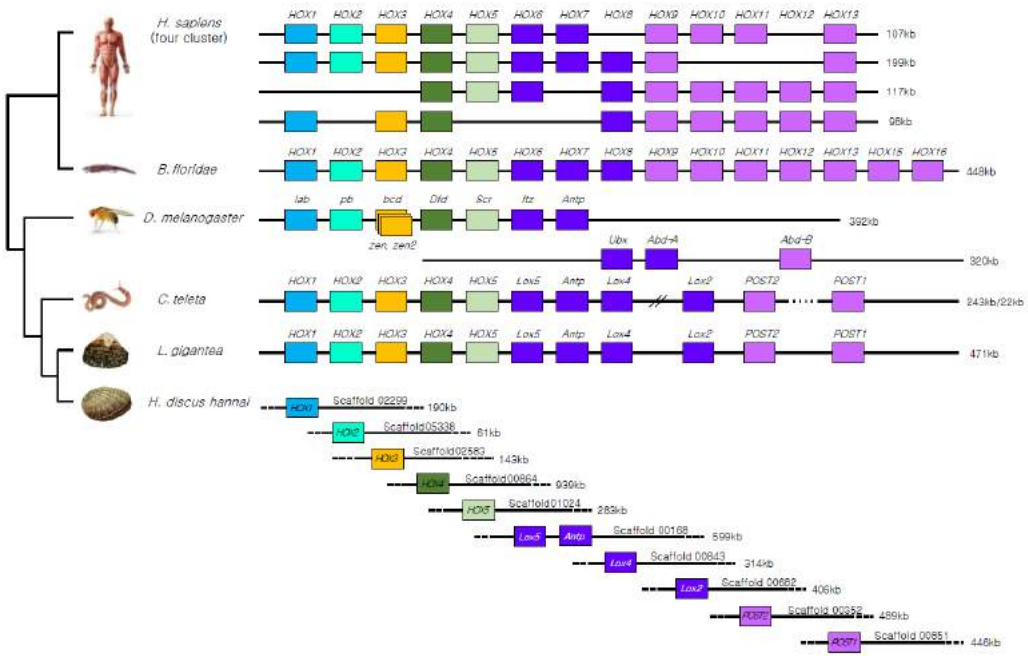



**Figure 3 - 11**. Atomized Hox gene cluster in Pacific abalone, *Haliotis discus hannai*.

At the top, the four Hox gene clusters of *H. sapiens* and the single Hox gene cluster of *B. floridae* are depicted. Hox gene clusters of *D. melanogaster* and *C. teleta* are split into two clusters. Hox cluster of *L.gignatea* is retained in the ancestral form. Note that, all Hox genes except *Lox5* and *Antp* were atomized in *H.discus hannai.* No additional homeo gene was identified within the same scaffold.

considered as an important genetic factor in characteristic phenotypes of the abalone such as enlarged muscular foot; the deviation from cluster preservation leads to a relaxation of regulatory constraints that enables various developmental modes.

**Repeat element and genome size of abalone**

The genome size of *H. discus hannai* was 1.86 Gb, and this is the biggest genome among known gastropods. It is 5.31 and 2.02 times larger than genomes size of *L.gigantea* (0.35 Gb) and *A.californica* (0.92 Gb) in the same Gastropoda class. In animals, the increase of genome size is commonly driven by transposable element, and this is a known genetic adaption mechanism to stressful environments(Chénais et al. 2012). Therefore, I conducted comparative analysis of repeat element against *L.gignatea*, a same marine gastropod with large genome size difference with that of *H.discus hannai*, to identify the reason for this large difference. Figure 3-12a shows the amount, proportion, and distribution of identified repeat element from two marine gastropods. The proportion of identified total repeat elements in *H.discus hannai* and *L.gigantea* is respectively 30.76% and 22.25% of genome size, and a total number of identified repeat elements in *H.discus hannai* genome is almost six times larger than that of *L.gigantea*. Such linear relationship

between genome size and the total number of repeat elements is consistent with a previous study(Kidwell 2002). The proportion, copy number and divergence of each mobile element were identified (Figure 3-13 ~ 3-16), for a deeper understanding of mobile elements in two species. From the comparison, a notable finding has been observed on mobile elements: DNA transposable element, a Class II transposable element, exists in diverse forms in both species; however, retrotransposon element, a Class I transposable element, is much more abundant in *H.discus hannai* genome than in *L.gigantea* genome. Especially, the number of a non-LTR retrotransposon called LINE Element was exceptionally high. Figure 3-12b illustrates the difference between the two species, using two signature mobile elements (*H.discus hannai*: LINE/I, DNA/TcMar-Tc1, *L.gigantea*: DNA/RC, DNA/Maverick) in each genome. DNA/RC and DNA/Maverick, two major mobile elements in *L.gigantea* genome, are observed in *H.discus* in somewhat similar distribution. On the other hand, the two signature mobile elements of *H.discus hannai* genome, LINE/I and DNA/TcMar-Tc1, are specifically abundant in *H.discus hannai* and seems to have expanded recently diverged compared to other elements. In sum, species specificity can be inferred from the distinctive patterns of repeat element expansion between the two species and the increased genome size of *H.discus hannai* may be associated with the non-LTR elements (especially

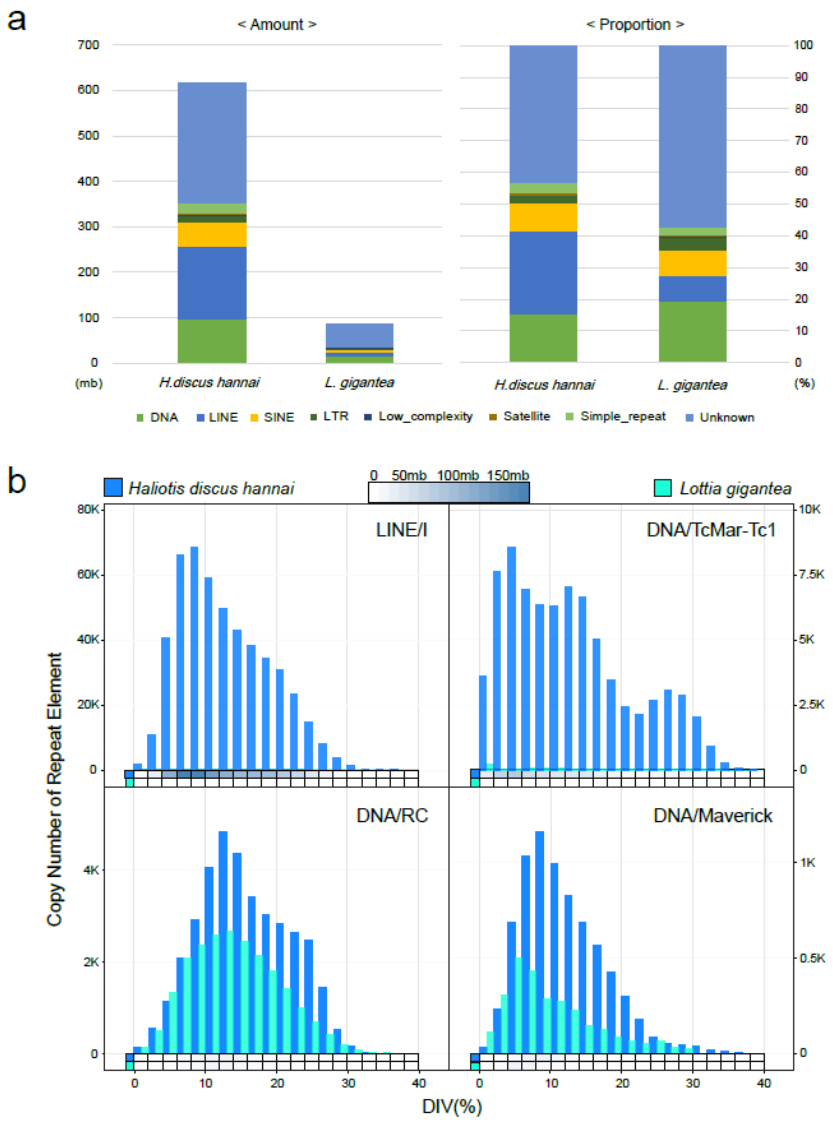



**Figure 3 - 12**. Repeat element information of *H.discus hannai* compared to *L. gigantean*.

a, Total amount and ratio of identified repeat element classified into 8 classes (DNA, LINE, SINE, LTR, Low complexity, Satellite, Simple repeat and Unknown) from each genome. b, Distribution of gene copy number of the two highly possessed repeat elements in each genome based on the divergence. Heat map indicates the total amount of repeat element divided into 20 levels based on the divergence.

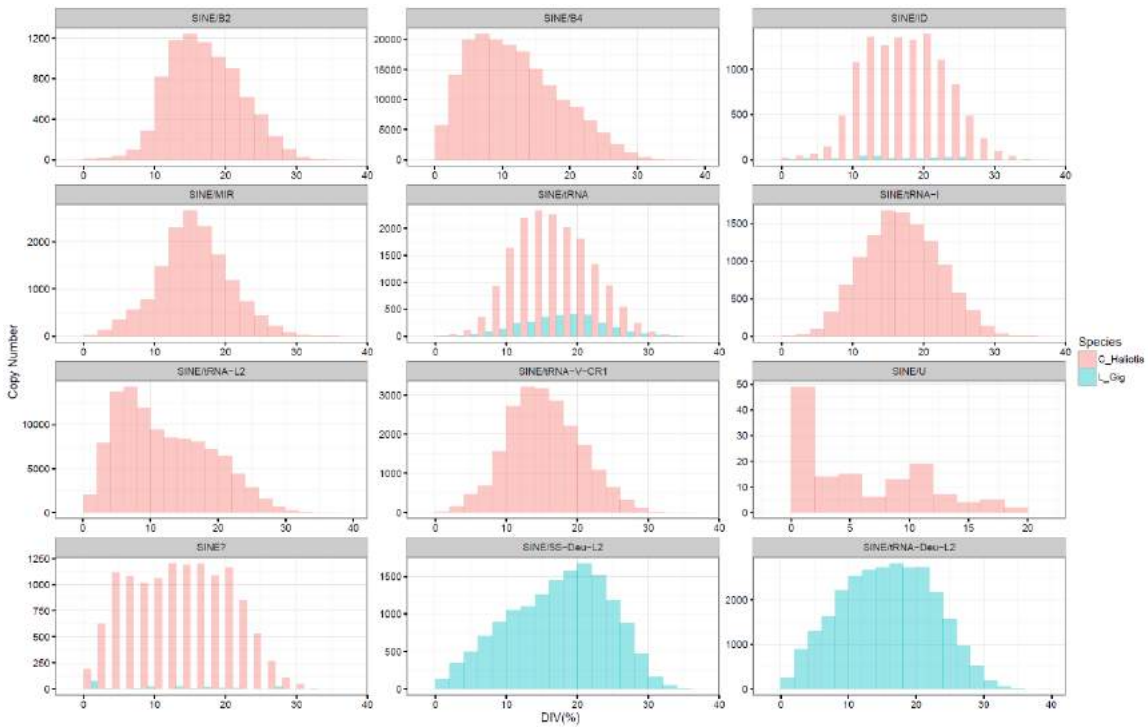



**Figure 3 - 13.** Comparison of SINE element distribution of *Haliotis discus hannai* and *Lottia gigantea*

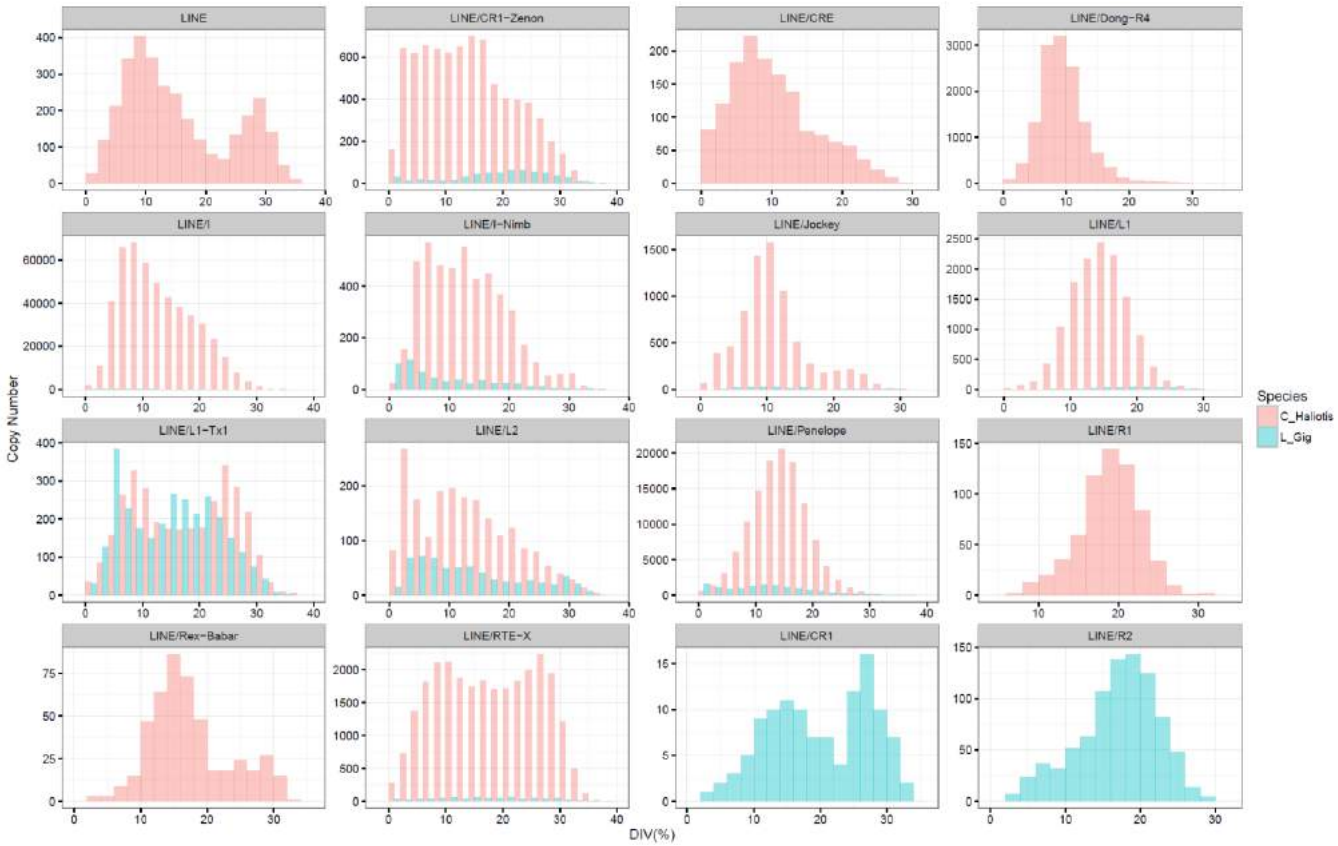



**Figure 3 - 14.** Comparison of LINE element distribution of *Haliotis discus hannai* and *Lottia gigantea.*

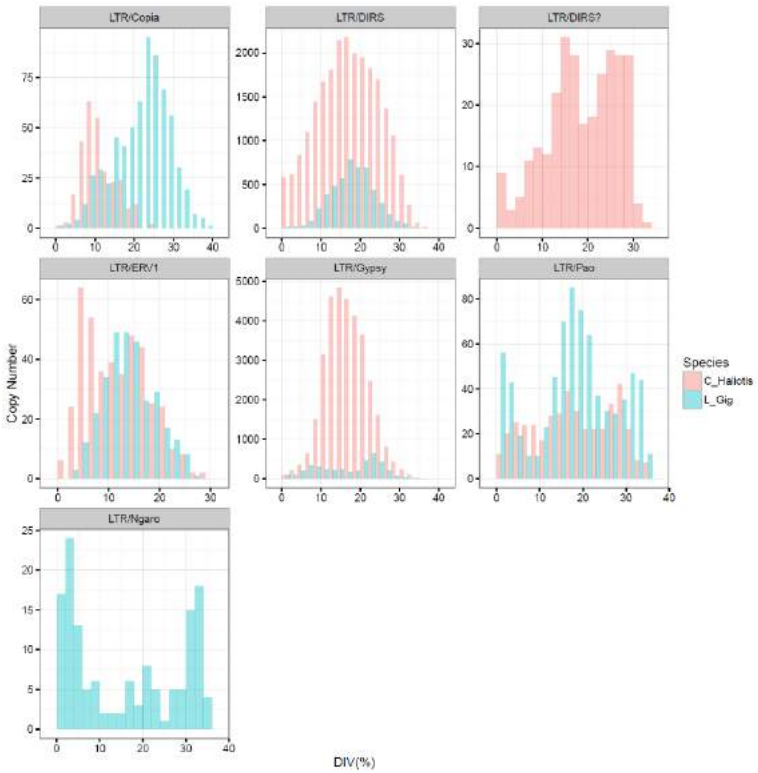



**Figure 3 - 15.** Comparison of LTR element distribution of *Haliotis discus hannai* and *Lottia gigantea*

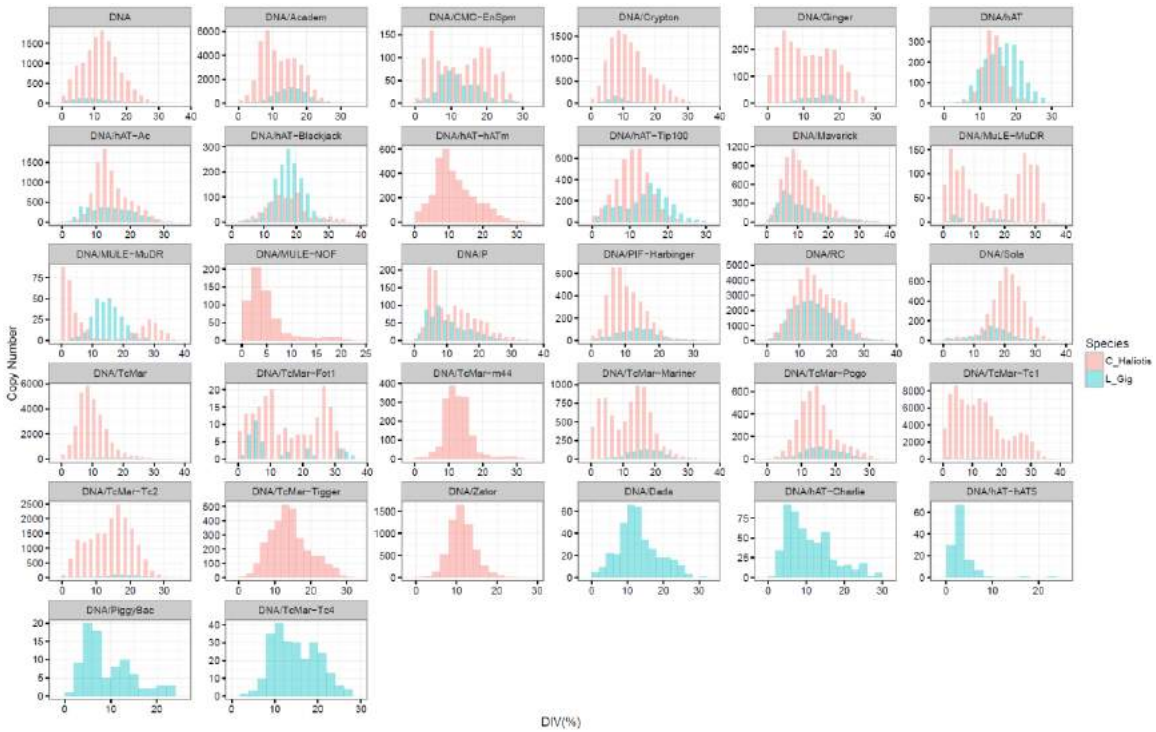



**Figure 3 - 16.** Comparison of DNA transposon element distribution of *Haliotis discus hannai* and *Lottia gigantea.*

LINE/I) contribution, in parallel to the human genome(Kidwell 2002). Limitations, however, in identifying the repeat elements through RepeatModeler and RepeatMasker, suggest polishing the rudimentary knowledge upon the link between repeat elements and evolution in Marine mollusk and gastropod genome; a fraction of repeat elements was masked as 'Unknown element.'

**Comparative Genome Analysis**

To study the abalone-specific evolutionary signatures that are preserved in their genome, comparative genome analysis against the genomes of 4 other marine animals was conducted (Figure 3-17). A total of 19,306 gene family clusters were constructed from the four species, and the *H.discus hannai* genome had 3,021 specific gene family clusters (Figure 3-17a). The *H.discus hannai* genome has more specific genes in comparison to the other four genomes (Figure 3-17b). In the results of Gene Family expansion/contraction analysis, abalone displayed 4,181 expanded and 7,211 reduced Gene Family clusters (Figure 3-17c). I summarized the species-specific genes and performed pathway analysis of the expanded clusters (Table 3-16), anticipating their importance in adaptation and evolution of *H.discus hannai*. I focused on the gene family expansions of two motor protein genes. The

abalone genome exhibits expansion in MHC (myosin heavy chain, striated muscle) gene and DNAH (Dynein heavy chain, axonemal) gene (Figure 3-17d). The *H.discus hannai* and *L.gigantea* each had 15 and 12 MHC genes related to striated muscle, respectively, which is almost fivefold that of *C.gigas* and *C.telleta* having 3 and 2, respectively. Additionally, the number of DNAH genes was 57 in *H.discus hannai* genome, which is more than twice those of *L.gigantea*, *C.gigas*, and *C.telleta* that are respectively 23, 25, and 26.

The foot is exceptionally enlarged and muscular in abalone, compared to other shellfish; this is thought to be a result of environmental adaption, to survive from assaults of predators and strong current. Also, abalone has the fastest moving speed among gastropods with unique crawling mechanism; it uses the combination of complex and rhythmic muscular contractions in the foot named *direct ditaxic* waves, which is the asynchronous wave of the foot muscle in a bisected manner(Miller 1974). GO pathway analysis about the expanded gene family showed that many expanded genes are related to muscle organization and functions such as M band, A band and sarcomere, etc. When looking at the gene families that comprise significant terms, the MHC genes dominantly appeared. MHC gene is the motor protein of muscle filaments and it provide motility via a power stroke mechanism fueled by the energy released from ATP hydrolysis(Burgess et al. 2003). Even though many genes are related

to muscle function, MHC gene is mainly responsible for contraction velocity of muscle fiber(Sweeney et al. 1988, Bottinelli et al. 1994). Previous study showed that individual muscle fiber simultaneously expressed various type of MHC and these degree of heterogeneity among individual muscle fibers can provide functional diversity for complex muscle movement(Stockdale and Miller 1987, Baldwin and Haddad 2001). MHC is also closely related to the backbone formation of the sarcomeric structure, which is designated as the contractile apparatus, in multiple stages of muscle development(Whalen et al. 1981, Harridge et al. 1996). Using comparative genome analysis between close marine animals, I identified that number of MHC genes related with striated muscle is commonly expanded in two marine gastropods (abalone and limpet) genome which have well developed muscular foot. Therefore, expanded MHC genes in abalone genome imply possibility for providing developmental and functional diversity in muscle fiber, and this may be associated with composition and complex movement of specific muscular foot of abalone.

DNAH gene causes sliding of microtubules-based cytoskeleton called the axonemes, which is another type of molecular machinery with dynein to generate force from ATP hydrolysis same as MHC(Lin et al. 2014). It is found only in those structures, and types of related DNAH gene are known to be

tissue-specific. In mammal, loss-of-function studies about DNAH gene showed that specific DNAH gene was related with the function of specific ciliary tissue(Pennarun et al. 1999, Neesen et al. 2001, Ibañez-Tallon et al. 2002). For example, mutation of DNAH7 affect the function of ciliary cell in respiratory tract and it is known to causes the disease named primary ciliary dyskinesia(Neesen et al. 2001). In abalone, DNAH is related to the structure of spermatozoa and egg during sexual maturation(Mendoza-Porras et al. 2014). Inferring from the expansion of DNAH gene families and previous studies about its function, our result indicates that expansion of DNAH gene may related to the functional diversity of microtubule-based structure such as cilia and flagella, and its related behaviors. In fact, motile cilia in molluscs can be expressed in different tissues throughout life cycle. For example, cilia on the surface of the gills of marine gastropods serve as sites for respiratory gas exchange(Taylor and Ragg 2005, Ragg and Taylor 2006) and it generate the water current under the shell holes in abalone through combination of ciliary beating movement(Ragg 2003, Taylor and Ragg 2005). Changes in gill structure have played a significant role in the diversification of gastropod group over their evolutionary history(Yonge , Lindberg and Ponder 2001). In addition, the epithelium of molluscs typically carries ciliated cells with glandular cells. Portela et al. have scanned the foot epithelium of the abalone

(*Haliotis tuberculate*) by light and electron microscopy. They revealed that the side foot epithelium is characterized by a microvillus border with a very scant presence of small ciliary tufts, but the sole foot epithelium bears a dense field of long cilia(Bravo Portela et al. 2012). The ciliated foot covered by a mucus layer is not only crucial for a range of functions including lubrication, locomotion, protection and adhesion(Davies and Hawkins 1998) but also plays an important ecological role in the community behavior(Hutchinson et al. 2007, Guo et al. 2009). Figure 3-17e shows the expression levels of DNAH genes in various abalone tissues. I identified that 57 DNAH genes in abalone genome are actually expressed, and expression patterns are tissue specific. This indicate that specific function of microtubule-based structures in various tissue may be associated with expansion and differential expression of DNAH genes. I believe that the gene family expansions of two major motor proteins (MHC and DNAH) in the abalone genome are possible explanations for remained vestiges of evolution in Haliotidae species. As of yet, the confirmatory analyses of such association study between motor protein gene expansion and their function associated target tissues are at a

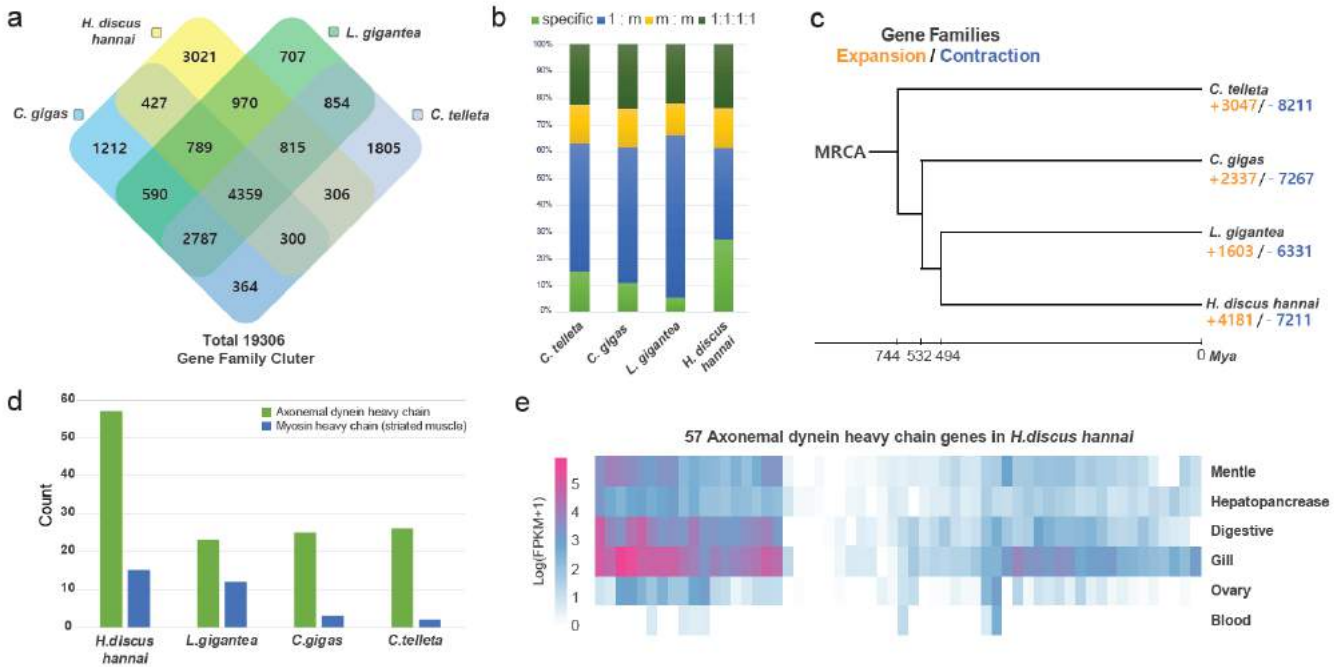



**Figure 3 - 17**. The result of comparative genome analysis of *H.discus hannai* with 3 close marine animals (*L.gigantea, C.gigas and C.telleta*).

a, Venn diagram shows the number of unique and shared gene families between 4 species. b, Proportion of genes in each genome classified into 4 categories. Specific indicates the unique genes. 1:m and m:m indicate one gene or multiple genes grouped into gene family with multiple genes from other species, respectively. 1:1:1:1 indicates 1:1 orthologous gene. c, Gene family expansion or contraction. The numbers indicate the amount of gene families that have been expanded (orange) or contracted (blue). d, Distribution of the number of axonemal dynein genes in 4 species. e, Expression of axonemal dynein genes in each organ tissue of *H.discus hannai***.**

rudimentary stage. I expect that further analysis about these genes will provide more information on Haliotidae evolution.

**Phylogeny and demographic inference of Pacific abalone**

To infer the historical changes in ancestral population size ($N_e$) of Pacific abalone, I employed the pairwise sequential Markovian coalescent (PSMC) method (Figure 3-18a). Our analysis uncovered the demographic trends of four Pacific abalone species from 1 million to 10,000 years ago. Based on the PSMC analysis, the results suggest that population size of four Pacific abalone species consistently increased before the last glacial period (approximately 110,000 to 12,000 years ago). Since the abalone are susceptible to water temperature, the elongated frigid conditions of the last glacial period acted as a bottleneck effect to decrease their population size steadily. In Late Glacial Maximum Period (13,000 to 10,000 years ago), which is also known as the beginning of the Modern Warm Period, three warm water species (*H.discus discus*, *H.madaka*, and *H.gigantea*) displayed an increase in population. In contrast, *H. discus hannai*, the cold water population, stayed at its low level and this indicates the difference in demographic history between the warm and cold water abalone species.

Abalones are distinguishable from other taxonomic groups for their very low spire and the small numbers of whorls, in Family level, yet lack phenotypic variation within their Genus level. Consequently, a recent study proposed a classification system for four Pacific abalones that are local to South Korea(Lee et al. 2014); it handles shell shape and solidity, the number of holes in the shell, etc. The shell attributes, however, are affected by ecological and nutritional factors that produce within-species variation, which increases the demand for molecular genetic markers. There are previous studies that used a few representative genetic markers to classify phylogenies in Pacific abalones, yet could not incorporate evolution history nor meet the classification standards of classic phylogeny(An et al. 2005). I, therefore, reconstructed whole MT genome, from whole genome sequencing data, and performed phylogenetic classification and divergence time estimation for the four Pacific abalones (Figure 3-18b). The overall topology divided into two groups: a single group consisting of 4 Pacific abalones (*H.discus hannai*, *H.disucs discus*, *H.madaka*, and *H.gigantea*) and the outgroup (*H.laevigata*, *H.rubra*, *H.tuberculata tuberculata*). Estimated divergence time compared to *H.discus hannai* are 1 million years ago, 2.36 mya, and 7.88 mya for *H.discus discus*, *H.madaka*, and *H.gigantea*, respectively. The results of a phylogenetic tree using whole MT genome, although different from those using a single gene

selected from MT genome, are consistent with that of the reported phylogenetic relationships. Previous studies that used single gene markers (i.e. COI gene) failed to reflect the known abalone genealogy, which suggests a single gene lacks resolution on classifying abalone species and subspecies(An et al. 2005). On the contrary, our demonstration of using the whole MT genome for classification shows promising results in abalone classification; as the first to analyze the phylogeny of Pacific abalones using whole MT genome, this approach is anticipated to resolve the limitations of classifying the abalone based on the shell attributes.

**Population structure of Pacific abalone**

To identify the population structure of Pacific abalone around South Korea, I collected 105 natural abalone samples consisting of *H.discus hannai*, *H.discus discus*, *H.gigantea,* and *H.rufenscens,* where *H.rufenscens* samples are an outgroup (Figure 3-19a). The genotype of each sample was identified using GBS and utilized in principal component analysis (PCA). The results display clear distinction of *H.discus hannai* samples, near South Korea, from each other by geographic differences, with minor exceptions (Figure 3-19b). The three seas near South Korea: Yellow Sea, South Sea, and East Sea, have varying environmental factors such as water temperature, depth, salinity, etc.

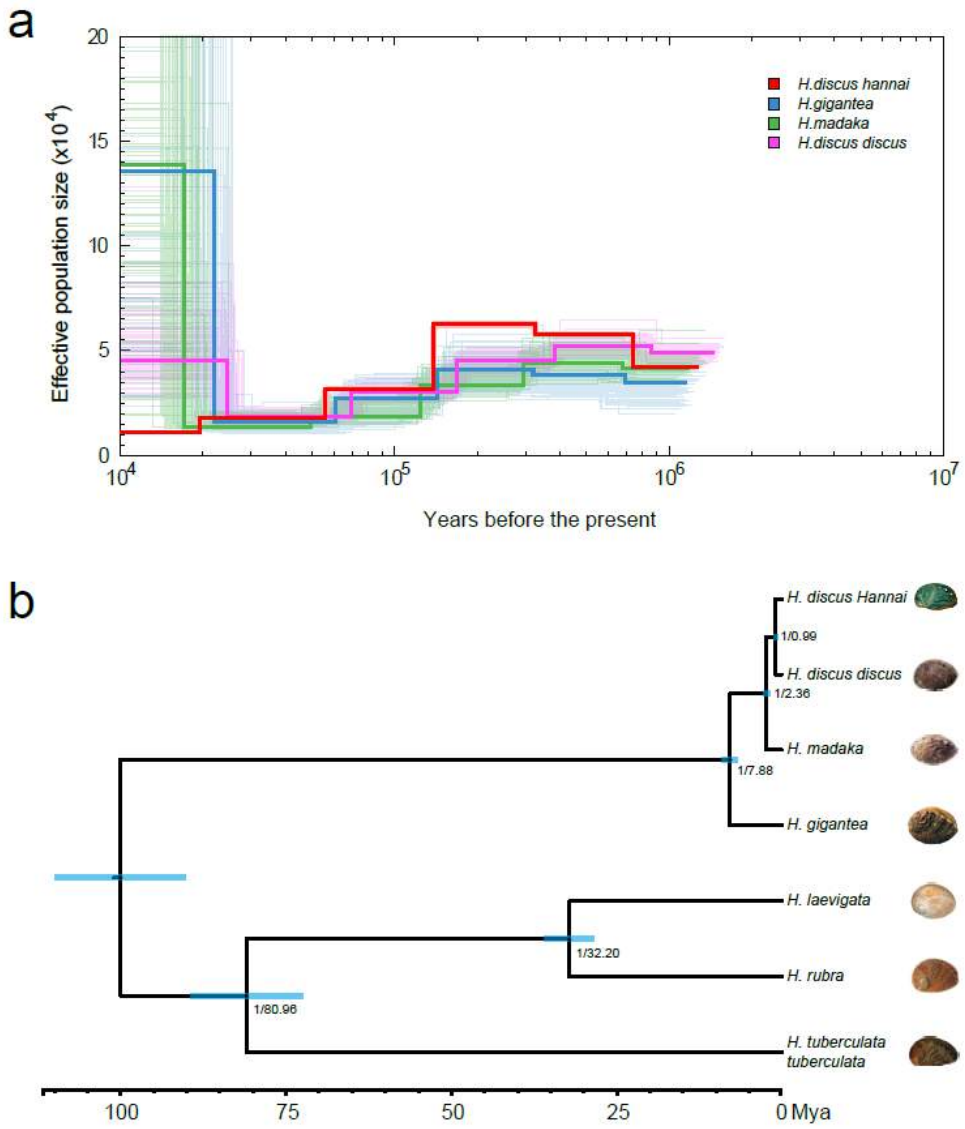



**Figure 3 - 18**. Demographic history and phylogeny of 4 Pacific abalones.

a, Pattern reconstruction of historical effective population size ($N_e$) for 4 Pacific abalone species. This pattern was inferred using pairwise sequential Markovian coalescent (PSMC) method. b, Phylogenetic tree of 4 Pacific abalone species with 3 outgroup species. Node value indicates the bootstrap value and estimated divergence time using BEAST2.

Since migration is unlikely and hard for abalones, the geographically different samples, within the same species, may contain life history and regional information in these abalones. To further understand the population structure and degree of admixture, in the population of Pacific abalone, I used STRUCTURE on randomly sampled SNPs (~61,795). I increased K value from 3 to 5, where K is the assumed number of ancestral population (Figure 3-19c). *H.rufenscens* samples, which are outgroup, displayed clear genetic homogeneity compared to others, and the Japanese samples exhibited a difference in admixture structure from that of the Korean *H.discus hannai* population. With K = 4, YEOSU and Japanese samples share relatively highest proportion of genome ancestry, out of the 3 *H.discus hannai* populations. I assume this has an association with current of sea and the distance being the shortest between YEOSU and Japan. Because larva of abalone can be moved followed the sea stream between Korea and Japan. The *H.discus hannai* population from TAEAN displayed higher homogeneity, in all K values, relative to the other regional populations, GOSEONG and YEOSU. This may be because the Yellow Sea is regionally isolated based on sea stream and has most distinctive environmental conditions out of the three seas; very low depth of water, high tidal range, and mud flat existence are some of its characteristics. Our results suggest that *H.discus hannai* reflect regional/environmental

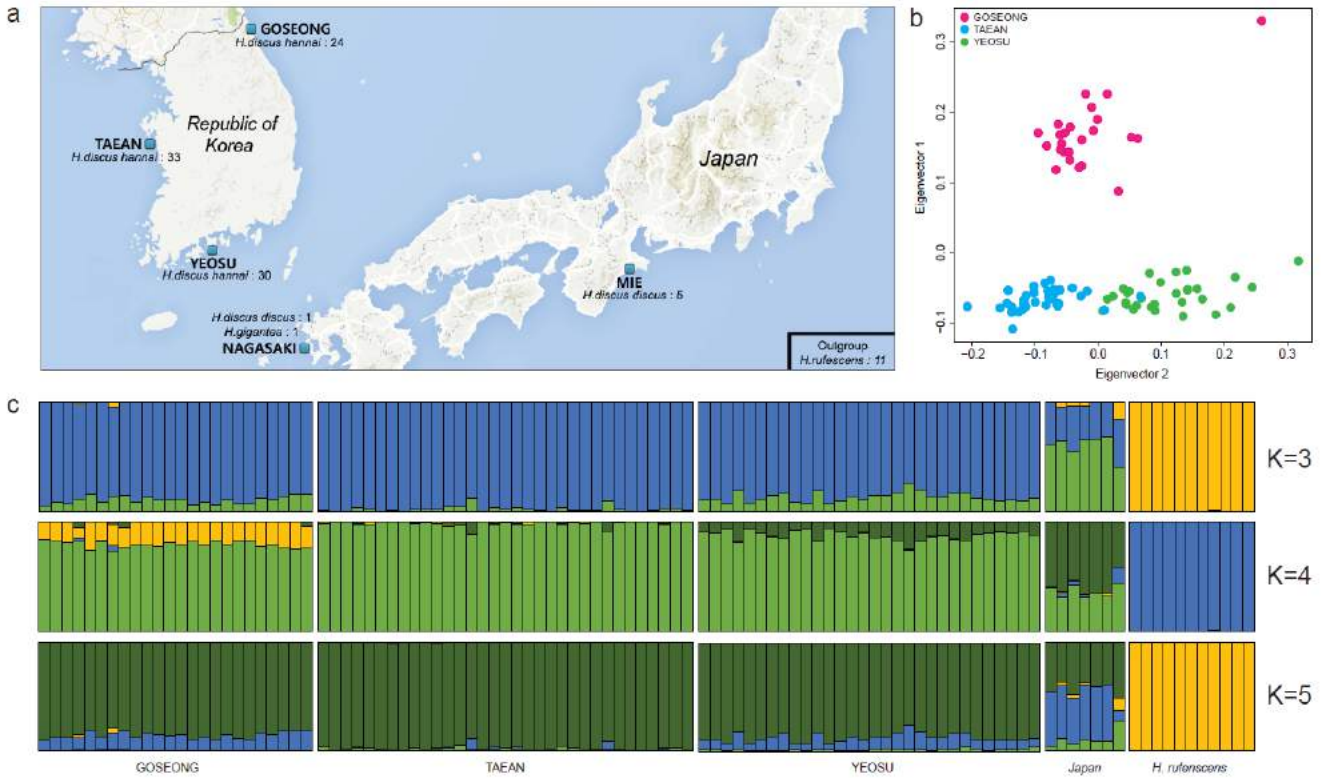



**Figure 3 - 19**. Geographical location, population structure, and relationships of Pacific abalone.

a, Sampling species, location and number of samples used for comparative population analysis. b, Principal component analysis of *H.discus hannai* from South Korea. c, Proportion of ancestry for each individuals assuming a different number of ancestral population (K=3,4, and 5). Colors in each individual represent the likelihood proportion of a genome assigned to a source population.

differences in its genome, and can be a good source of research in terms of ecological and evolutionary studies.

**Microbial community and meta-transcriptome analysis of abalone visceral extract**

A previous study showed that the visceral extract of *H.discus hannai* had tumor suppression effect on breast and lung cancer(Lee et al. 2010, Suleria et al. 2015). However, such tumor suppression mechanism is yet to be explained. To identify the potential clue and material of visceral extract against carcinogenic burden, I employed microbial community and meta-transcriptome analysis. Alpha-diversity value of visceral extract was 9.812, and rarefaction curve is shown in Figure 3-9. Figure 3-20a shows the microbial community in the visceral extract of *H.discus hannai*. Among identified microbes in phylum level, Fusobacteria and Tenericutes were dominant microbes and two dominant phyla occupied 74.5% of the total microbial community (Figure 3-10). In genus level, the majority of Fusobacteria phylum was Illyobacter and of Tenericutes phylum was Mycoplasma. Through meta-transcriptome analysis, variously expressed genes were reconstructed, and related functions were identified. Among related pathways of genes, I identified that enzymes related to the ceramide

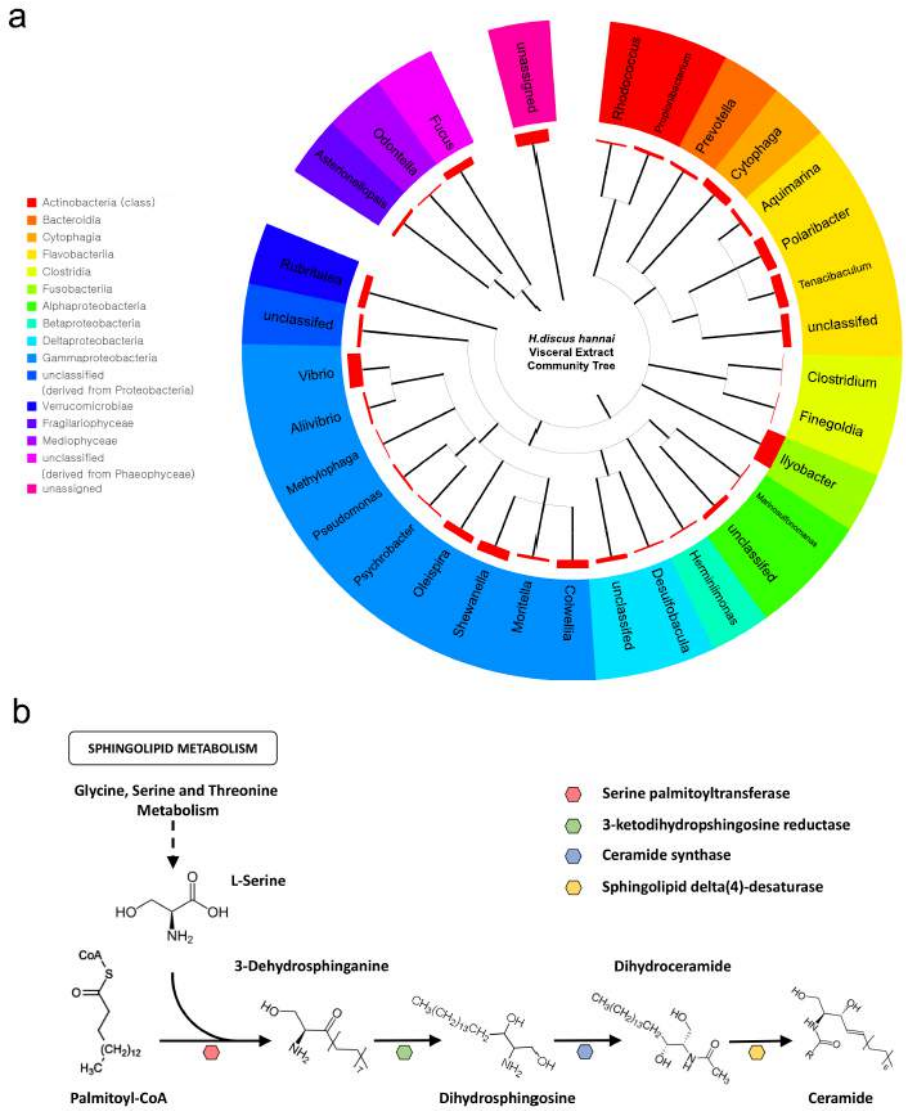



**Figure 3 - 20**. Microbial community and activated ceramide synthesis pathway in abalone visceral extract.

a, Community tree of abalone's visceral extract using 16s rRNA in genus level. Red bar indicates the abundance of each genus. b, Ceramide synthesis pathway and identified foreign enzymes expressed in abalone's visceral extract.

synthesis in sphingolipid metabolism of KEGG pathway are expressed in the visceral extract of abalone (Figure 3-20b). To identify the origin of that activated pathway, nucleotide sequences of expressed enzymes were compared with those of *H.discus hannai* genes. Enzyme sequences related to ceramide synthesis in abalone viscera contain novel sequences that are not from *H.discus hannai*.

Ceramide is a family of waxy lipid molecules and is composed of sphingosine and a fatty acid. It is known to have an apoptosis-inducing effect and received attention as a tumor suppressor lipid. Numerous studies have demonstrated tumor suppression effect of ceramide(Charles et al. 2001, Kurinna et al. 2004, Ogretmen and Hannun 2004, Reynolds et al. 2004, Morad and Cabot 2013). In meta-transcriptome analysis, I identified various enzyme sequences of foreign origin that are related to the ceramide synthesis process in abalone viscera. One of the main genera in the microbial community of abalone viscera was Illyobacter. It is an anaerobic microbe with various fermentation pathways(Stieb and Schink 1984), and it is known to synthesize ceramide from precursor substrates using sphingolipid metabolism pathway. These results indicate that ceramide in abalone viscera is a product of gut microorganisms, and previous studies showed that abalone's visceral extract contains a considerable amount of ceramide(Matsubara et al. 1990) and its

metabolites(Toshiko and Akira 1982). Therefore, I expect that ceramide from gut microbiome can be one of the potential anti-cancer materials in abalone's visceral extract. However, such anti-carcinogenic properties do not depend on the ceramide alone; there are reports on the abalone's visceral extract down regulates the expression of Cox2 gene, a gene which is over-expressed in cancer cells as a defense mechanism against ceramide apoptosis-inducing effect. By the way, abalone's visceral extract contains various bioactive compounds such as phloroglucinol, fucoxanthin, and fucoidan that originate from brown seaweed, one of the main feeds of *H.discus hannai(Pádua et al. 2015)*. Another study illustrated brown algal polyphenol suppression of the cyclooxygenase-2 (COX-2) expression(Hwang et al. 2006, Kim and Chojnacka 2015) and cell proliferation in tumor cells(Jung et al. 2009, Taylor 2011, Thomas and Kim 2011).

Overall, I can hypothesize that tumor suppression of abalone's visceral extract is a compound effect of ceramide from gut microbiome which induces apoptosis in tumor cells and the bioactive compound from brown seaweed, which suppress the cell proliferation via down-regulation of the COX-2 gene. However, owing to the conflicting and inconsistent nature of bioactive compound mechanisms, illustrating the underlying mechanisms of such tumor suppressing effects is demanding. Despite these limitations, our analysis is the

first to perform microbial community and meta-transcriptome analysis on the abalone's visceral extract to identify potential candidates of tumor suppression effect. Through analyses above, I propose a notable candidate for cancer suppressing mechanisms of abalone's visceral extract and hope our study to provide more insight into further studies in this area.

### ***H.discus hannai*** **genome database**

I built the *H.discus hannai* genome database for sharing our results and data to research community. *H.discus hannai* genome database is available in http://gdb.insilicogen.com/haliotis_discus_hannai/.

# Chapter 4. Mitochondrial Genome Assembly: Effective complete mitochondrial genome assembly pipeline for unprecedented species

## 4.1 Introduction

The mitochondria can be found in almost eukaryotic organism. It is usually described as “the powerhouse of the cell” because it generates adenosine triphosphate(ATP), used as a source of chemical energy. This is semiautonomous organelle which functions in aging process, apoptosis, anti-HIV drugs and cancers(Kroemer and Reed 2000). And the genome sequence of this is one of the most widely used markers in molecular phylogenetics(Gray et al. 1999). Because mitochondrial genome sequence is one of commonly existing homologous sequences in every eukaryote except very rare case.

Mitochondrial genome consists of conserved and variable regions, and these regions can be useful for identifying evolutional relationship in various scopes of divergence. Based on the Motoo Kimura’s molecular clock theory(Kimura 1980), closely related organisms have high degree of similarity, while organisms distantly related show a dissimilarity pattern. Because mutations can be accumulated over time, this can be used as molecular clock for divergence time. However, constructing complete mitochondrial genome sequence for unprecedented is challenging. Genome size of mitochondrion varies from approximately 14,000bp to 20,000bp, and this cannot be sequenced at a time using conventional sequencing. Therefore, primer walking

was essential for building the complete genome sequence of mitochondrion using conventional sequencing. This method extends the initial DNA sequence using repeatedly designed custom primer. However, designing proper primers and extending initial fragmented sequence take quite amount of times and experimental efforts. And sequencing cost is also expensive compared to the current NGS system. To resolve these limitations, specific algorithms based on NGS platforms were developed. For example, MITOBIM(Hahn et al. 2013) is the most widely used *de novo* assembly program for reconstructing complete mitochondrial genome using Illumina short read. This program uses the algorithm called "baiting and iterative mapping approach". Baiting is identifying initial sequence based on the closely related organism and iterative mapping is used for iterative extension using the overlap consensus layout assembly method. However, when I tested this program for unprecedented species which have no reference sequence of closely related species, several attempts for constructing complete sequence of mitochondrial genome was ended in failure. Various reasons were expected for this result and optimized pipelines were needed to solve this problem.

Therefore, in this study, I built the complete mitochondrial genome assembly pipeline using Illumina sequencing platform including experimental process and components of this pipeline is based on the benefits of various NGS study.

I applied this pipeline to one of unprecedented species named *Menathais tuberosa* and the complete mitochondrial genome sequence was successfully assembled. Constructed pipeline has advantages such as cost, time and success rate over conventional method. I expect this constructed pipeline contribute to constructing more complete MT genome and provide more opportunity for molecular phylogenetic studies via expanding research resources.

## 4.2 Materials and Methods

### Library Construction and Sequencing

In general, usual whole genome sequencing data contains quite amount of mitochondrial genome sequence. For obtaining high coverage mitochondrial genome sequence, however, mitochondria have to be enriched before DNA isolation for sequencing library. MT DNA was isolated using Mitochondrial DNA Isolation Kit (BioVision, CA, USA) and, Illumina PCR free library preparation kit was used for sequencing library construction with insert size 500bp according to the manufacturer's protocol. To generate the raw data, 300 cycle paried-end sequencing using Miseq was conducted

### Data Processing and Seed Sequence Mining

The quality of the whole genome sequencing reads from all samples was checked using FastQC (Andrews 2010) and rawdata quality control process was conducted using Trimmomatic(Bolger et al. 2014) with the parameter(ILLUMINACIP:TruSeq3-PE-2.fa:2:30:10, MINLEN:250). Paired-end reads retained after basic quality control process were merged into single read using FLASH(Magoč and Salzberg 2011).

To identify the proper seed sequence for mitochondiral genome assembly, whole available mitochondrial genome sequence of various species were downloaded from NCBI refseq database(Pruitt et al. 2007). Paired-end read were aligned to mitochondrial genome sequence database with local alignment option using Bowtie2(Langmead and Salzberg 2012). Alignment file modification and coverage calculation was conducted using Samtools(Li et al. 2009), and candidate seed sequence was isolated based on the continuous coverage (>min. 600bp) and mapping depth of uniquely mapping reads.

**Assembly and Finalize MT genome**

MITOBIM v1.8 was employed for MT genome assembly with identified candidate seed. Merged single reads from pair-end reads using FLASH were assembled with default option except iteration number changed to 100. Seed sequence extension was iteratively conducted until extended sequence cannot elongated. Length of assembled sequence was checked with reference mitochondrial genome sequence of the closest neighbor. To check the quality of assembled sequence, paired-end read mapped using Bowtie2 with no-mixed option. Selection of assembly was conducted based on the evaluation using length of assembled sequence, mapping coverage and depth, and mapped distance between paired-end read among candidate assemblies. Assembly

polishing step was conducted to reduce the assembly errors of MITOBIM assembly. For this, variant calling process iteratively conducted using Genome Analysis Toolkit. The “UnifiedGenotyper” of GATK(McKenna et al. 2010) was used for calling candidate single nucleotide variants (SNVs) and InDel. To avoid possible false positive variants, argument “VariantFiltration” of the same software was adopted with the following options: 1) SNVs with a phred-scaled quality score of less than 30 were filtered; 2) SNVs with QD(unfiltered depth of non-reference samples; low scores are indicative of false positives and artifacts) <5 were filtered; 3) SNVs with FS (phred-scaled P value using Fisher’s exact test) >200 were filtered as FS represents variation on either the forward or the reverse strand, which are indicative of false-positive calls; 4) SNVs with MQ0(the number of reads which have mapping quality zero) >4 or MQ0/DP(proportion of mapping quality zero reads over total depth) >0.1 were filtered to remove uncertain calls; 5) more than 3 SNVs within 10bp window were filtered. Identified variants were altered to assembled genome sequence used for the reference genome in re-sequencing process and this step was repeatedly conducted until identified variants minimized. After polishing the assembled sequence, sequence circularization was conducted using Circulator(Hunt et al. 2015). Orientation and direction of assembled sequence

was manually adjusted using BioEdit(Hall 1999). After sequence construction, gene annotation was conducted using MITOS(Bernt et al. 2013).

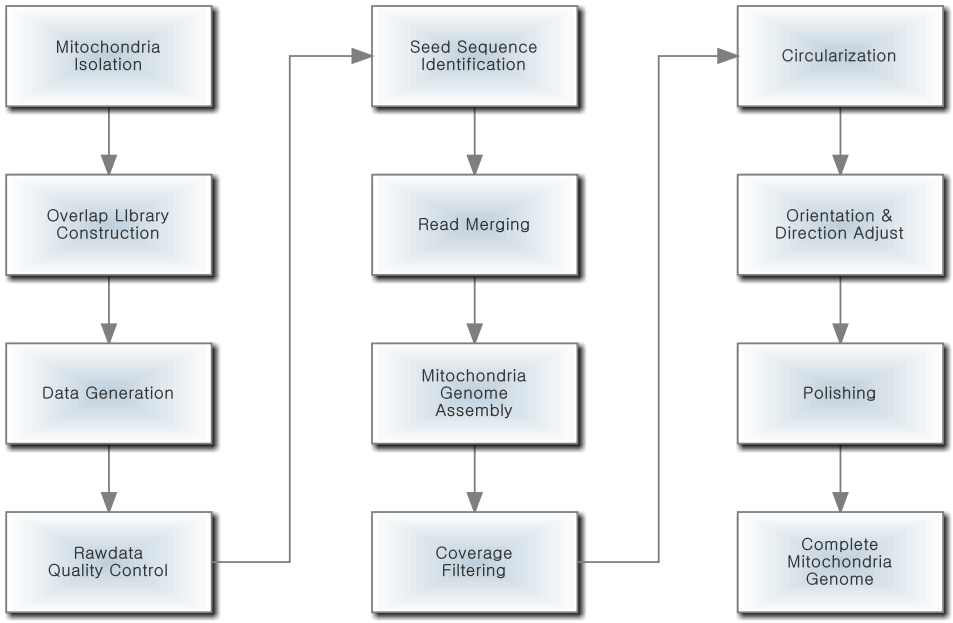



**Figure 4 - 1.** Constructed complete mitochondrial genome assembly pipeline.

## 4.3 Results and Discussion

### Constructed Pipeline for complete MT genome assembly

Figure 4-1 shows the constructed complete mitochondrial genome assembly pipeline. In this pipeline, I applied 3 experimental steps in sequence data generation. First, mitochondria were enriched before library construction. In whole genome sequencing data contains quite amount of mitochondrial genome. However, in various test using whole genome sequencing data, complete mitochondrial genome cannot be constructed because seed sequence cannot be extended enough. The reason of this is based on the lacking sequence coverage. Therefore, more sequence information of mitochondria is needed. Mitochondrial genome enrichment before sequencing can provide enough reads for the constructing compete mitochondrial genome. Next, PCR free library construction kit was used for data generation. It is known that there are various AT rich regions in mitochondrial genome sequence. Sequencing of these AT rich or GC rich regions cannot be properly conducted because of PCR bias. Illumina sequencing platform, including conventional sanger sequencing, rely on the PCR amplification. Therefore, library construction using PCR free library kit can provide exhaustive sequences without unamplified regions. Last, fragment size selection was conducted based on the same criteria used in metagenome community analysis. An assembler used in this pipeline is based

on the OLC(Overlap Layout Consensus) algorithm and read length is one of the most important considerations. To obtain more read length, paired-end reads are merged into single read. This is normally used for capturing V3-V4 regions in 16S rRNA in metagenome community analysis. Merged read can provide approximately 450-550bp reads and this can provide more possibilities of extension for complete mitochondrial genome assembly. In assembly step, seed sequence was identified from all available mitochondrial genome sequence in NCBI refseq database. Previous method uses conserved sequences such as COI (Cytochrome C Oxidase I) of closest neighbor. However, when I compared the COI sequence of the closest neighbor with the seed sequence from whole mitochondrial genome sequence database based on the sequence similarity, two sequences showed many differences. This indicated that closest neighbor did not always provide a proper seed sequence for extension. Therefore, read alignment using Bowtie2 local alignment to refseq mitochondrial genome database can provide custom seed sequences regardless of evolutional distance. And the polishing step used in Pacbio RS system have been applied, too. This method is usually used for reducing errors in assembly because read from Pacbio system contains quite amount of errors (>10%). However, this can be also applied for Illumina data and make the assembly result more accurate.

**Application for real data**

Mitochondrial genome sequences of *Menathais tuberosa* were generated to verify the constructed pipeline (Table 4-1). In the seed sequence identification, 1,004bp of *Menathais bimaculata* COI gene sequence(GI|378924595) was selected for backbone seed sequence. Assembly pipeline was carried out with selected backbone sequence and complete mitochondrial genome sequence was successfully built. The length of assembled mitochondrial genome for 15,294bp which is similar to the other species belonging to the same family such as *R. clavigera* (15,285 bp), *R. venosa* (15,272 bp) and *B. brandaris* (15,380 bp). AT contents was 65.9%, which is lower than *R. clavigera* (66.2%), *R. venosa* (69.0%) and *B. brandaris* (66.4%). The mitochondrial genome is composed of 13 protein-coding genes, 2 ribosomal RNA genes, and 22 tRNA genes with 4 overlapping regions between 2 and 26 bp in length. Gene information and genetic map are showed in Figure 4-2.

**Advantages of constructed pipeline**

My result using *Menathais tuberosa* data showed that constructed pipeline can be useful for assembly of complete mitochondrial genome sequence of unprecedented species. And constructed pipeline has advantages compared

**Table 4 - 1.** Summary statistics for the *Mennathais tuberosa* mitochondrial genome.

| **Assembled Mitochondrial Genome** | |
|---|---|
| Size(1n) | 15,864 bp |
| GC level | 34.07% |
| Number of A's | 4,509 bp |
| Number of C's | 2,615 bp |
| Number of G's | 2,790 bp |
| Number of T's | 5,950 bp |

| Name | Start | Stop | Strand | Length | Structure |
|---|---|---|---|---|---|
| cox3 | 384 | 1136 | + | 753 | |
| trnK(ttt) | 1177 | 1244 | + | 68 | svg ps |
| trnA(tgc) | 1246 | 1312 | + | 67 | svg ps |
| trnR(tcg) | 1321 | 1389 | + | 69 | svg ps |
| trnN(gtt) | 1396 | 1462 | + | 67 | svg ps |
| trnI(gat) | 1482 | 1548 | + | 67 | svg ps |
| nad3 | 1552 | 1902 | + | 351 | |
| trnS1(gct) | 1919 | 1986 | + | 68 | svg ps |
| nad2 | 2014 | 2940 | + | 927 | |
| cox1 | 3050 | 4561 | + | 1512 | |
| cox2 | 4608 | 5276 | + | 669 | |
| trnD(gtc) | 5293 | 5360 | + | 68 | svg ps |
| atp8 | 5362 | 5517 | + | 156 | |
| atp6 | 5539 | 6213 | + | 675 | |
| trnM(cat) | 6263 | 6329 | - | 67 | svg ps |
| trnY(gta) | 6333 | 6398 | - | 66 | svg ps |
| trnC(gca) | 6400 | 6462 | - | 63 | svg ps |
| trnW(tca) | 6463 | 6528 | - | 66 | svg ps |
| trnQ(ttg) | 6530 | 6591 | - | 62 | svg ps |
| trnG(tcc) | 6607 | 6674 | - | 68 | svg ps |
| trnE(ttc) | 6675 | 6739 | - | 65 | svg ps |
| rrnS | 6815 | 7698 | + | 884 | svg ps |
| trnV(tac) | 7696 | 7763 | + | 68 | svg ps |
| rrnL | 7742 | 9139 | + | 1398 | svg ps |
| trnL1(tag) | 9114 | 9183 | + | 70 | svg ps |
| trnL2(taa) | 9193 | 9261 | + | 69 | svg ps |
| nad1 | 9274 | 10167 | + | 894 | |
| trnP(tgg) | 10211 | 10279 | + | 69 | svg ps |
| nad6 | 10293 | 10739 | + | 447 | |
| cob | 10802 | 11896 | + | 1095 | |
| trnS2(tga) | 11937 | 12001 | + | 65 | svg ps |
| trnT(tgt) | 12002 | 12069 | - | 68 | svg ps |
| nad4l | 12080 | 12373 | + | 294 | |
| nad4 | 12388 | 13731 | + | 1344 | |
| trnH(gtg) | 13753 | 13818 | + | 66 | svg ps |
| nad5 | 13846 | 15519 | + | 1674 | |
| trnF(gaa) | 15533 | 15600 | + | 68 | svg ps |

tRNA gene rRNA gene protein coding gene

**Figure 4 - 2.** Genetic map information of *M. tuberosa* mitochondrial genome**.**

to conventional Sanger sequencing. The first advantage is cost. In normal case, sequencing cost is under 1/3 compared to primer walking method. Using NGS platforms with the constructed pipeline, mitochondrial genome sequence can be effectively assembled with low cost. The second advantage is analysis time. Primer walking is time consuming process and designing proper primers fitted to experimental condition is challenging. At least one month is need for completing assembly using primer walking. However, constructed pipeline in this study can be carried out within one week. In addition, this method can be also conducted in parallel using the multiplexing features of NGS platforms. Therefore, many mitochondrial genomes can be constructed using one sequencer. The third advantage is success rate of assembly. This constructed pipeline was tested for various species in other studies and successfully built the all complete mitochondria genome sequence(An et al. 2016, Karagozlu et al. 2016, Karagozlu et al. 2016, Sung et al. 2016, Sung et al. 2016, Sung et al. 2016). However, success rate of conventional method is much lower than this method. Because conventional sequencing method rely on the PCR reaction and PCR bias make AT rich region cannot be properly analyzed. But, PCR free library used in this pipeline provide more possibility regardless of AT content in mitochondrial genome. As such, mitochondrial genome assembly pipeline constructed in this study can resolve the limitations of previous method and I

expect that this pipeline provide more opportunity for researchers who are interested in mitochondrial genome of unprecedented species.

# Chapter 5. Microsatellite Marker Build: Development of microsatellite markers and data analysis for population identification

## 5.1 Introduction

With the advancement of genotyping technologies, population genetics analysis can provide an effective way for research in these days. Various types of genomics variants such as SNP(Single Nucleotide Polymorphism), InDel(Insertion and Deletion) and SV(Structural Variation) can be leveraged. A microsatellite is a type of genetic variant which have certain DNA motifs. It ranges in length from 2-5 base pairs and repeated 5-50 times. It is commonly located in non-coding region of genome, but they also can be located in regulatory and coding region. In non-coding regions, mutations in microsatellite can be accumulated because it is comparably free from possible links with the specific function. And it also has a higher mutation rate other types of variants, so it can provide useful genetic diversity information. In addition, microsatellite is a multi-allelic variant. It can provide more genotype information compared to SNP which have only 3 different type of genotype(Jehle and Arntzen 2002). Based on these characteristics, a microsatellite is one of power analytic tools for DNA fingerprinting and identification. In case of human, this genetic variation is widely used in forensic field and parentage test. However, for unprecedented species, building microsatellite marker is challenging because of absence of reference genome. Without reference genome, locations of microsatellites and flanking

sequences for primer design cannot be properly conducted. However, building reference level genome rely on the huge amount of research fund and times even though sequencing price continuously decrease. Therefore, building microsatellite marker for unprecedented species with efficient and economical way demands. For this, some previous study showed that identifying microsatellite can be identified from whole genome shotgun sequencing data(Castoe et al. 2012). However, just profiling types of microsatellite maker is not enough for further study. It's because markers need flanking genome sequence for primer design. Benefits using microsatellite marker compared to whole genome sequencing are low cost and high information. PCR is one of the most commonly used methods and identified markers for downstream analysis must have properly designed PCR primer.

In this study, I made a pipeline for constructing microsatellite marker of unprecedented species based on *de novo* assembly. Whole genome sequence data of *Antheraea yamamai* with low coverage was used and constructed microsatellite markers were validated in practice. Validation of constructed marker showed that constructed pipeline can be helpful for developing microsatellite markers and researches about unprecedented species. Using this method, I expect that many researchers, interested in unprecedented species, can conduct their studies more easily in an economical manner.

## 5.2 Materials and Methods

### Library construction and data generation

Total 12, 3 samples for each region, samples of *Antherothera yamamai* was collected from 4 regions (Pocheon, Yeoungwol, Hamyang and Seokwuipo) (Figure 5-1). Genomic DNA was extracted from muscle tissue using a DNeasy Blood and Tissue kit (QIAGEN, Hilden, Germany) following the manufacturer's protocol. The quality of genomic DNA was checked on 1% agarose gels with a spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Sequencing library of one sample from Seokwuipo was constructed using TrueSeqnano DNA Library Kit (Illumina, San Diego, CA) following the manufacturer's protocol. Sequencing of the prepared library was conducted using Miseq (Illumina) with 250 bp paired-end reads.

### Data processing and Microsatellite Identification

The quality of the whole genome sequencing reads was checked using FastQC (Andrews 2010) and rawdata quality control process was conducted using Trimmomatic(Bolger et al. 2014) with the parameter(ILLUMINACIP:TruSeq3-PE-2.fa:2:30:10, MINLEN:200).


**Figure 5 - 1.** Sampling locations of 12 *A. yamamai* samples used in this study.

Sequencing errors were corrected using error correction module of ALLPATHS-LG(Gnerre et al. 2011). After error correction, genome assembly was conducted using IDBA_UD(Peng et al. 2012) with pre-correction option. To identify reliable assembly, short reads were mapped to assembled sequences using Bowtie2(Langmead and Salzberg 2012), and mapping depth and coverage was calculated using Samtools(Li et al. 2009). Only assembled contigs with depth > 10, coverage > 90% were used for microsatellite marker identification. Locations of microsatellites were identified using Msatcommander(Faircloth 2008). Primers for microsatellite marker was built using Primer3(Rozen and Skaletsky 1999) with MIN_SIZE=20, MAX_SIZE=24, MIN_TM=48, MAX_TM=52, OPT_TM=50 options. Among constructed primer sets, 10 locations were selected based on the microsatellite types, read depth and product size.

**Genotyping and Data analysis**

One of each primer pair was labeled with 6 FAM fluorescent dyes (Yueet al., 2000). I conducted PCR in a25 μℓ reaction in a ABI 2720 Thermo cycler. Conducted PCR program as follow: a first denaturation step at 95℃ for 3 minutes, followed by 30 cycles at 94℃ for 30 seconds, annealing temperature

at 50℃ for 30 seconds and 72 for 1 minute, and a final extension for 5 minutes at 72℃. Each reaction contained concentration 30 ng of DNA, 1 × PCR buffer [50mM KCl, 10mMTris-HCl (pH 8.8), 150 nM KCl, 1.5mM MgCl2], 2.5mM of dNTPs, 200 nM of each primer, and 1U of Pure Speed PFU DNA polymerase (Smarteome, KOREA). After PCR, I mixed 0.2 ㎕ of PCR product with 9.8 ㎕ Hi-Di Formamide (Applied Biosystems, USA) and 0.2 ㎕ Liz-500 standard size (Applied Biosystems, USA). The samples were denatured at 95℃ for 5 minutes and placed on ice. ABI 3730xl (Applied Biosystems, USA) was employed for separation of mixed products and analysis. Allele size calling and genotyping were carried out using the GeneMapper® Software v4.1 (Applied Biosystems, USA).

Observed heterozygosity, expected heterozygosity, PIC(Polymorphic Information Content), the allelic and genotypic frequencies, phylogenetic tree using Neighbor Joining Method were calculated using PowerMarker ver. 3.25(Liu and Muse 2005).

## 5.3 Results and Discussion

### Constructed Pipeline for microsatellite building of unprecedented species

Figure 5-2 shows the constructed pipeline for microsatellite marker build of unprecedented species. Basically, it relies on the *de novo* assembly of whole genome sequence. After assembly, it identifies microsatellite in assembled contigs and designing primer using flanking sequence around microsatellite region. Specificity of designed primer was tested using assembled genome and this can reduce the possibility of wrong amplification. Previous study what I conducted for minke whale(Park et al. 2015) using low coverage data showed that even low coverage whole genome sequencing data can be assembled to contigs enough to identifying genomic contents. Therefore, researchers can conduct their population genetic studies for unprecedented species without huge amount of research funds for the genome project.

### Microsatellite marker build for *Antheraea Yamamai*

I generated approximately 14X coverage genome sequencing data for *Antheraea yamamai* using Miseq platform (Table 5-1). Using constructed pipeline, initial contigs of *Antheraea yamamai* were assembled for

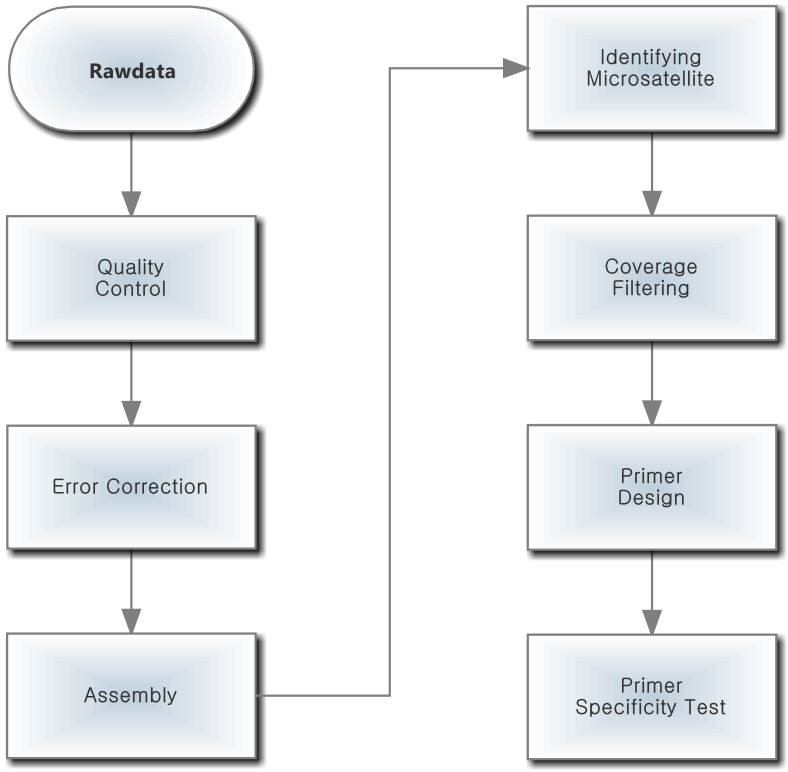



**Figure 5 - 2.** Microsatellite marker build pipeline for unprecedented species.

**Table 5 - 1**. Summary statistics of generated whole genome shotgun sequencing data.

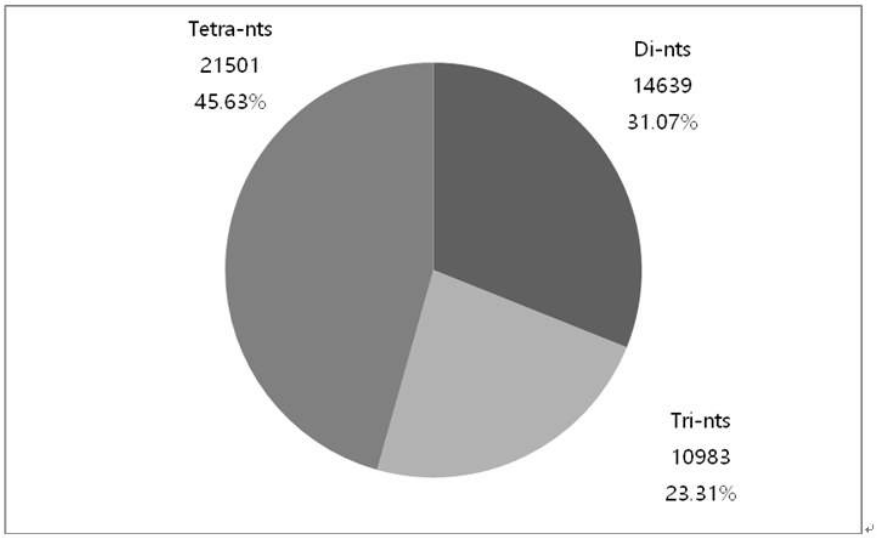
| Library Name | Library Type | Insert Size | Platform | Read Length | No. Read | Total bp |
|---|---|---|---|---|---|---|
| AY | Paired-end | 500 | Miseq | 250 | 40,670,816 | 10,208,374,816 |

downstream analysis. Table 5-2 shows the assembly result of *A. yamamai* genome. Assembled genome size was approximately 0.71Gb and GC level was 34.78%. N50 length was 2,750bp and average contig length was 752.7bp. Assembled genome size is matched with known *A. yamamai* genome and N50 length indicated that assembled genome had been assembled enough to identify flanking sequence of microsatellite region. Microsatellite regions in assembled genome were profiled using Msatcommander (Figure 5-3). Figure 5-3A shows the proportion of 3 different types of tandem repeats. The proportion of Di nucleotide and tri nucleotide was 31.07% and 23.31%, respectively. Tetra nucleotide took the largest portion (45.63%). Figure 5-3B shows the distribution of 4 most frequent repeat motifs. AT, AAT and CTGT repeat were the most frequent motif for each type of repeat, respectively. Among the identified microsatellite, 10 markers were selected based on the read remapping coverage, uniqueness of designed primer, and type and motif of tandem repeat. Table 5-3 shows the constructed set of 10 microsatellite markers. Annealing temperature varies in 49 to 53 and product size ranges from 123bp to 263bp.

**Table 5 - 2**. Summary statistics for the *Antheraea yamamai* draft genome.

| Assembled Genome | |
|---|---|
| Size(1n) | 0.71 Gb |
| GC level | 34.78% |
| No. scaffolds | 949,962 |
| N50 of scaffolds (bp) | 2,750 |
| Longest(shortest) scaffolds (bp) | 74,727 (124) |
| Average scaffold Length (bp) | 752.7 |

**A.**




**B.**

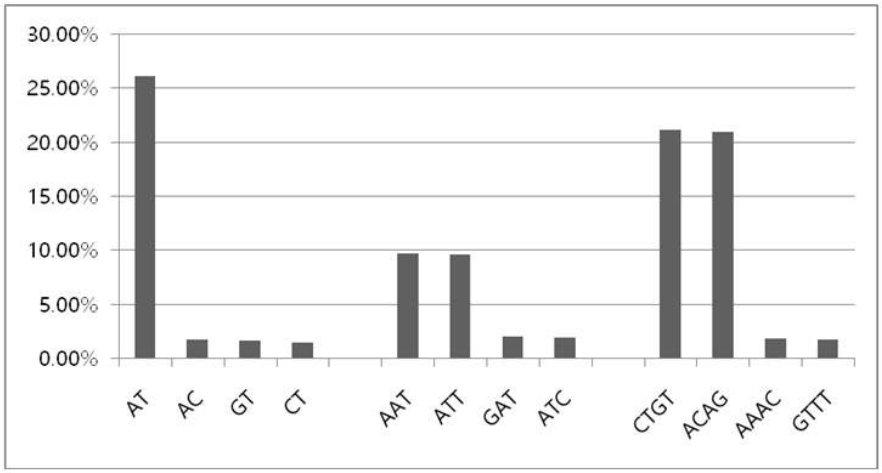



**Figure 5 - 3**. Proportion and distribution of identified microsatellite in *A. yamamai* genome. (A) Distribution of di ~ tetra-nucleotide microsatellite on contigs with a minimum 10X coverage. (B) Distribution of most frequent repeat motifs with minimum 10X coverage.

**Table 5 - 3.** The set of 10 microsatellite markers developed in *Antheraea yamamai.*

| Marker name | Repeat Motifs | primer sequence(5'-3') | annealing(℃) | Size Range(bp) | GenBank Accession No. |
|---|---|---|---|---|---|
| 10545 | (AAT)12 | GGGTCTACAAAGAAATCTTAT | 52 | 263 | KJ735682 |
| | | GTCGTCTTGAAATGTTTTTA | 49 | | |
| 10908 | (AAT)15 | CTGAGTACCAGATTAAAACA | 51 | 123 | KJ735684 |
| | | ATACCCTGGTTAAAAACAAT | 49 | | |
| 12519 | (ATT)12 | GCATTCATTAGACAAATACCA | 52 | 272 | KJ735688 |
| | | CGTCATGGAGAGAATATATC | 53 | | |
| 57897 | (AC)12 | TCGATCTGGTTTATCTTTTTAA | 50 | 103 | KJ735719 |
| | | ATGTCTGGAATTAACTCATC | 51 | | |
| 74763 | (GAT)15 | GATATACGAGTAGAGGAAGA | 53 | 204 | KM582134 |
| | | TAAAAACCCAAGACATACAA | 49 | | |
| 8575 | (AT)20 | GTTCAGTCGGTCATATTAAT | 51 | 223 | KM582137 |
| | | CGCTAAAAACAATAACAGAA | 49 | | |
| 47017 | (AG)46 | CTAAATTTGTCTTTACCAGC | 51 | 234 | KM582149 |
| | | TATATCGCCAGTTATATTGG | 51 | | |
| 30993 | (ATC)33 | GCAGTTTATTGAATAATGCT | 49 | 177 | KM582156 |
| | | ATATAGCCCATAGTCTTCTT | 51 | | |
| 123729 | (GGT)30 | TGGAATTCACCAAGATAATT | 49 | 209 | KM582158 |
| | | TATATAGCCTATAGCCTTCC | 53 | | |
| 50112 | (AGT)30 | CACGACTTTAGTTAGTACTT | 51 | 245 | KM582159 |
| | | ATAAGGGTTATACCAAGCTA | 51 | | |

**Genotyping and Phylogenetic Analysis**

Genotyping was conducted using constructed 10 microsatellite markers. Table 5-4 shows the summary of genotyping result for 12 *A. yamamai* samples. 10 microsatellite markers were successfully genotyped without missing. In case of marker 10545 and 10908, there were only two types of alleles which is same with the SNP. These locations cannot provide more information compared to SNP, so these markers have to be replaced. Marker 47017 showed the highest number of alleles and genotypes, this marker contains many information compared to other markers. However, too many genotypes in one marker have difficulties in genotyping process. The number of genotype was 2-11 and the mean was 5. Expected heterozygosity was 0.1528-0.8681 and observed heterozygosity was 0.1667-1.0000. The PIC (Polymorphism in Content) value, which is used as the criteria of polymorphism for markers, was 0.1411-0.9581. PIC value of all markers except two markers, 10545 and 10908, had PIC value over 0.5 and selected markers were thought to contain information for downstream analysis. Phylogenetic tree using neighbor joining method was constructed based on the calculated using shared allelic method (Figure 5-4). Seokwuipo is the city of Jeju island and population of *A. yamamai* is thought to be isolated from other samples from land population. In the phylogenetic analysis, Seokwuipo samples showed close shared allelic distance and they

**Table 5 - 4.** Summary statistic of 10 microsatellite markers in *Antheraea yamamai.*

| Marker | Major allele Frequency | Genotype No. | Sample Size | No. of obs. | Allele No. | Availability[a] | $H_e$[b] | $H_o$[c] | PIC[d] |
|---|---|---|---|---|---|---|---|---|---|
| 10545 | 0.9167 | 2 | 12 | 12 | 2 | 1 | 0.1528 | 0.1667 | 0.1411 |
| 10908 | 0.5417 | 3 | 12 | 12 | 2 | 1 | 0.4965 | 0.7500 | 0.3733 |
| 57897 | 0.4583 | 3 | 12 | 12 | 5 | 1 | 0.6806 | 1 | 0.6284 |
| 47017 | 0.2917 | 11 | 12 | 12 | 13 | 1 | 0.8681 | 0.9167 | 0.8581 |
| 12519 | 0.5000 | 5 | 12 | 12 | 3 | 1 | 0.6111 | 0.6667 | 0.5355 |
| 74763 | 0.5833 | 6 | 12 | 12 | 5 | 1 | 0.6076 | 0.4167 | 0.5705 |
| 50112 | 0.7083 | 4 | 12 | 12 | 4 | 1 | 0.4618 | 0.5833 | 0.4247 |
| 123729 | 0.3333 | 9 | 12 | 12 | 6 | 1 | 0.7569 | 0.7500 | 0.7185 |
| 8575 | 0.2917 | 7 | 12 | 12 | 6 | 1 | 0.7917 | 0.8333 | 0.7607 |
| 30993 | 0.3750 | 9 | 12 | 12 | 8 | 1 | 0.7708 | 0.8333 | 0.7422 |
| Mean | 0.5000 | 5 | 12 | 12 | 5.4000 | 1 | 0.6198 | 0.6917 | 0.5753 |

[a]Availability is defined as 1-*Obs* / n, where *Obs* is the number of observations and n is the number of individuals sampled.

[b]Gene diversity, often referred to as expected heterozygosity, is defined as the probability that two randomly chosen alleles from the population are different.

[c]Heterozygosity is simple the proportion of heterozygous individuals in the population.

[d]A closely related diversity measure is the polymorphism information contents (PIC).

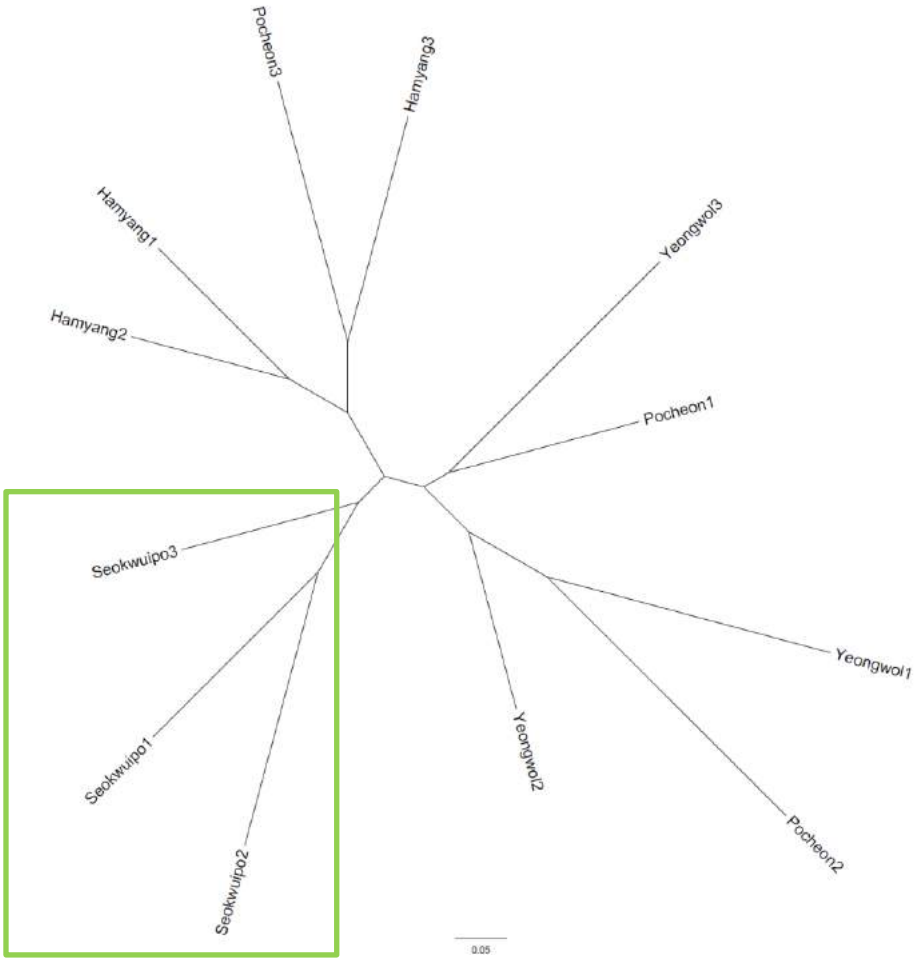



**Figure 5 - 4.** Neighbor-joining tree for genetic relationship among 4 populations of *Antheraea yamamai.*

Pairwise genetic distances between each sample were calculated based on the shared allelic methods(Chakraborty and Jin 1993) using the PowerMarker.

were grouped together. However, samples from land population were mixed. This indicate that there might be migrations of individuals between interior population and selected makers cannot provide enough information for regional classification. However, results of genotyping and phylogenetic analysis showed the possibility that constructed pipeline can be used for race classification. In practical application, national institute of biological resource developed microsatellite marker set for classification of local *Dorcus hopei* from exotic variety using this microsatellite build pipeline. I hope this developed pipeline provide more efficient ways for researchers and public officers whose works are related with unprecedented species.

# Chapter 6. Unaligned Read Assembly: Identifying Hanwoo specific genome variation using unaligned read assembly

## 6.1 Introduction

Reference genomes have provided efficient ways for analyzing an individual genome in many ways. Especially, the analysis method called “re-sequencing” was widely used for NGS data to supplement the weakness of short read length from Illumina platforms. One of the most representative project was 1,000 human genome project(Siva 2008). Researchers identified SNPs from different human population and compared it. It is based on the assumption that the difference between the reference genome and an individual genome is very small. This means that target individual must have close evolutional relationships with the species used for reference genome. Analysis process of this method is conducting alignment of short reads from NGS platform to the reference genome, and then identifying differences between them. From this analysis process, the types of variants such as SNP, small InDel is mainly focused. The reason of this is originated from short read length of current sequencing platforms and it is related with low specificity of reads. Correct mapping of reads is the most important process for reducing false-positive variant. For this, only reads with few base different can be allowed for mapping to increase the specificity of an individual read. However, this makes difficulties at the identifying the large difference, such as structural variant, between the reference genome and an individual sample. To solve these

limitations, various algorithms and programs were developed. Most of programs for detection of structural variation such as BreakDancer(Chen et al. 2009), PIndel(Ye et al. 2009), and GenomeSTRiP(Handsaker et al. 2011) are based on the 4 algorithms known as “Read pair”, “Read Depth”, “Split Read”, and “Assembly”(Medvedev et al. 2009). “Read pair” algorithm uses the distance information between two reads of paired-end sequencing system. For example, if paired reads were mapped to the reference genome with significantly different distance compared to insert size of constructed library, I can identify that there is a deletion in the sample genome. Using “Read depth”, we can identify CNVs (Copy Number Variation) with significantly increased or decreased read mapping depth based on the assumption of random shearing in library construction. However, this algorithm has weakness in identifying the novel sequence insertion because novel sequence cannot be mapped to the reference genome. “Split Read” algorithm uses local alignment. This algorithm checks the split alignment of single read from local alignment and I can detect structural variant signal from junction alignment. However, as mentioned above, read length of Illumina sequencing system is short. Even though divided read can be mapped to the reference genome, specificity of read mapping cannot be guaranteed with more short reads. This problem can be easily solved if I use long reads over 2kbp such as Pacbio

system. However, Pacbio system have low throughput with high error rate compared to Illumina system, so it is not proper for an individual genome analysis to in these days. “Assembly” algorithm is an efficient way to identify structural variations because it can identify not only locations but also sequence information of variants. Especially, unaligned read assembly, conducting assembly of unaligned read from short read alignment, is known to be one of efficient ways for solving the limitations of re-sequencing. However, unaligned read assembly is challenging area because of the underlying assumptions and algorithms of *de novo* assembly of NGS data.

To solve this practical limitation, I built the unaligned read assembly pipelines for Illumina whole genome shotgun sequencing data. Constructed pipeline was successfully applied to previous studies and I also applied this method to the new Hanwoo population data. Gaps in the reference genome and candidate structural variants might be related with breeding process of Hanwoo were identified. I hope that constructed pipeline can provide more valuable variant information and contribute to the resolving of limitations in usual re-sequencing process.

## 6.2 Materials and Methods

### Unaligned read assembly and annotation

Total 136 Hanwoo genome data was used in this analysis. Table shows the used data information in this study. The quality of the whole genome sequencing reads was checked using FastQC (Andrews 2010) and rawdata quality control process was conducted using Trimmomatic(Bolger et al. 2014) with the parameter(ILLUMINACIP:TruSeq3-PE-2.fa:2:30:10, MINLEN:125). Sequencing errors were corrected using error correction module of ALLPATHS-LG(Gnerre et al. 2011). After error correction, genome assembly was conducted using IDBA_UD(Peng et al. 2012) with pre-correction option. Filtered reads were aligned to reference cattle genome UMD3.1 using Bowtie2(Langmead and Salzberg 2012) with default option. Unaligned reads were isolated using Samtools(Li et al. 2009) based on the FLAG and unaligned reads were aligned to assembled contigs. Mapping depth of unaligned reads were calculated using Samtools and common regions between 136 Hanwoo individuals were identified using in-house python script. Mapping depth under 20X were filtered for downstream analysis. For the gene annotation of unaligned read assembly, augustus(Stanke et al. 2006) was employed for gene prediction using mammal parameter. Function of predicted genes were identified using UMD 3.1 reference genome annotation and

InterproScan5(Jones et al. 2014). Gene ontology pathway analysis was conducted using the ClueGO(Bindea et al. 2009) add-on of Cytoscape(Shannon et al. 2003).

## 6.3 Results and Discussion

### Constructed Pipeline and Unaligned Read Assembly

Constructed pipeline showed in Figure 6-1. Same as above studies, this method also relies on the whole genome assembly. Constructed pipeline conduct assembly using not just isolated unaligned reads but whole genome data. It's because of the error correction algorithm of *de novo* assembly. Many NGS read assemblers using de bruijn graph method and this is conducted based on the k-mers from raw reads. Unlike OLC(overlap layout consensus) method, de bruijn graph using k-mer is weak to sequencing error because k-mer graph cannot be directly continued if there is one base different in k-mer. This make the graph complicated and assembly cannot be conducted successfully. Therefore, error correction is one of the most important steps in assembly process based on *de bruijn* graph method. However, error correction algorithms of genome assembly were based on the assumption that sequencing data is whole genome shotgun sequencing data. When I conduct whole genome sequencing, I assume that generated data contains whole genomic regions with equal depth. Based on this assumption, I can also assume that every k-mer have to be appeared at least average sequencing depth. For example, if 10X coverage sequencing data was converted to k-mer, each k-

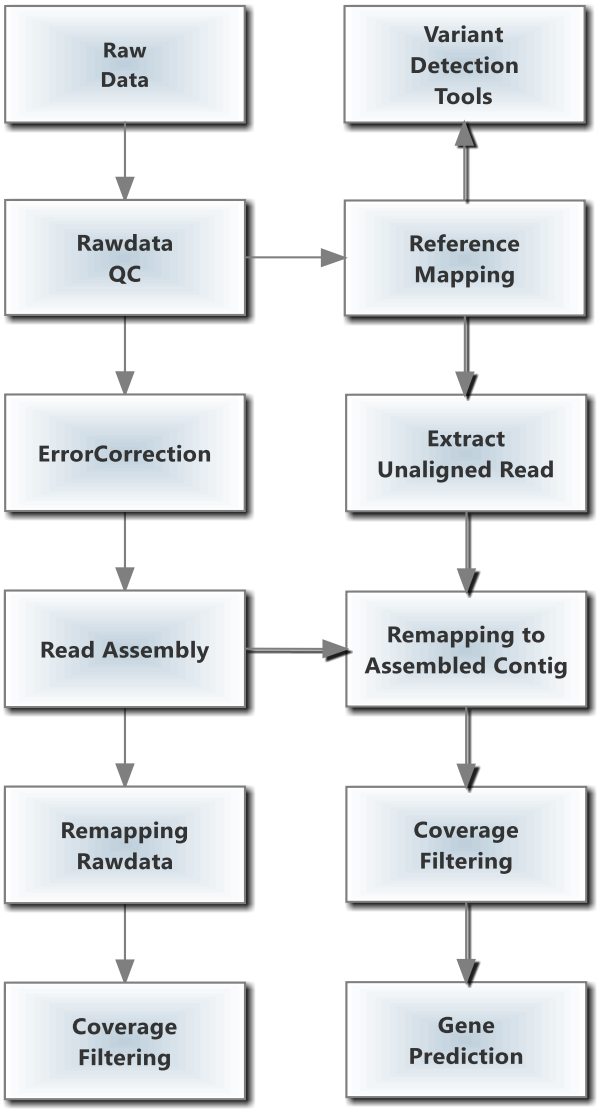



**Figure 6 - 1.** Workflow of constructed unalinged read assembly pipeline.

mer have to be appeared at least 10 times. In this situation, k-mer appeared only one or two time could be identified as a sequencing error. Most of error correction modules of *de novo* assemblers using Illumina NGS platform rely on this method. Therefore, if I conduct error correction for assembly just using unaligned reads, error correction process cannot be properly conducted. To remove this limitation, constructed pipeline uses the alternative way. Unaligned reads are not solely assembled but mapped to whole genome assembly. Unaligned read assembly result is identified from mapping depth of remmaping. This method has benefits that flanking region sequences of unaligned read assembly can be identified. Assembled contigs only using unaligned read don't contain same region of the reference genome. However, using this method, researchers can identify the genomic coordinate based on reference genome using flanking region sequence.

136 Hanwoo samples were aligned to reference genome using Bowtie2. Average alignment rate was 95.22% (Figure 6-2) and approximately 5% of reads were unaligned. And one of samples from 136 Hanwoo population was assembled (Table). Assembled genome size was approximately 2.98Gb and GC level was 42.5%. N50 length was 10,127bp and average contig length was 661.45bp. Same as previous study conducted for *A. yamamai* genome,

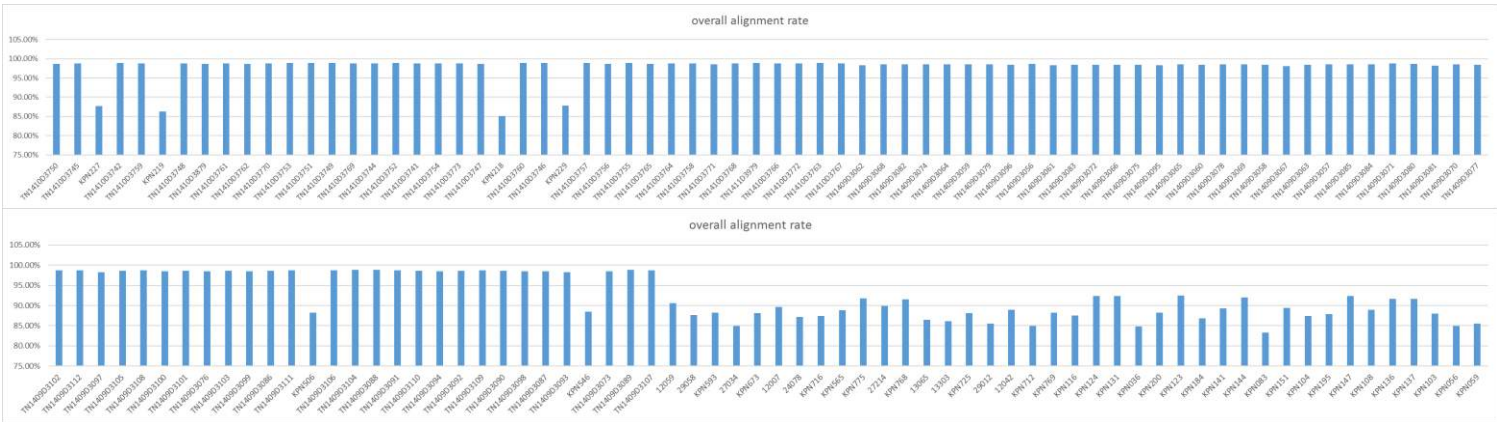



**Figure 6 - 2.** Distribution of read alignment of 136 Hanwoo samples using Bowtie2.

**Table 6 - 1.** Summary statistics for the Hanwoo contig assembly.

| Assembled Genome | |
|---|---|
| Size(1n) | 2.98 Gb |
| GC level | 42.5% |
| No. scaffolds | 4,517,620 |
| N50 of scaffolds (bp) | 10,127 |
| N bases in scaffolds (%) | 7,398,050 (0.25%) |
| Longest(shortest) scaffolds (bp) | 136,625(100) |
| Average scaffold Length (bp) | 661.45 |

assembled genome size is matched with reference genome. After remapping unaligned read from 136 Hanwoo samples to assembled sequence, commonly mapped regions from all sample which indicate common sequence of unaligned read assembly in Hanwoo population were isolated.

**Genes in Unaligned Assembly and its Function**

After gene prediction and functional annotation, 179 genes were identified in unaligned reads. To identify the related functions of these genes, GO (gene ontology) pathway analysis was conducted using ClueGO. Table 6-2 and Figure 6-3 shows the enriched pathways and network of enriched pathways(p-value<0.05). Network analysis on enriched GO term showed that identified genes were related with positive regulation of mitotic nuclear division, skeletal muscle tissue regeneration, skeletal muscle cell differentiation, substantia nigra development, regulation of heart rate and phospholipid dephosphorylation. Hanwoo is the native cattle breed of Korea and it has evolved to be meat type cattle over the period of time. Strong breeding program was initiated from 1980's and the result of breeding program might be recorded in Hanwoo genome. GO pathway analysis showed that identified genes located in unaligned read assembly is related with muscle development. It is a reasonable result because breeding program of Hanwoo was mainly

**Table 6 - 2.** Enriched GO terms and related genes using ClueGO analysis. (p-value < 0.05)

| GO ID | GO Term | P - Value | Associated Genes |
|---|---|---|---|
| GO:0002027 | regulation of heart rate | 0.0057 | EDN1, KCNJ2, MC3R, SRI |
| GO:0021762 | substantia nigra development | 0.01 | CNP, COX6B1, SYPL2 |
| GO:0045931 | positive regulation of mitotic cell cycle | 0.0029 | ANXA1, EDN1, EREG, NSMCE2, UBE2C |
| GO:0051785 | positive regulation of nuclear division | 0.0013 | EDN1, EREG, NSMCE2, UBE2C |
| GO:1901989 | positive regulation of cell cycle phase transition | 0.013 | ANXA1, NSMCE2, UBE2C |
| GO:0045840 | positive regulation of mitotic nuclear division | 0.00058 | EDN1, EREG, NSMCE2, UBE2C |
| GO:1901992 | positive regulation of mitotic cell cycle phase transition | 0.01 | ANXA1, NSMCE2, UBE2C |
| GO:0035914 | skeletal muscle cell differentiation | 0.018 | BCL9, PAX5, ZNF689 |
| GO:0042246 | tissue regeneration | 0.013 | ANXA1, BCL9, SOX15 |
| GO:0043403 | skeletal muscle tissue regeneration | 0.0021 | ANXA1, BCL9, SOX15 |
| GO:0046839 | phospholipid dephosphorylation | 0.0032 | INPP5K, PLPPR4, PTRH2 |

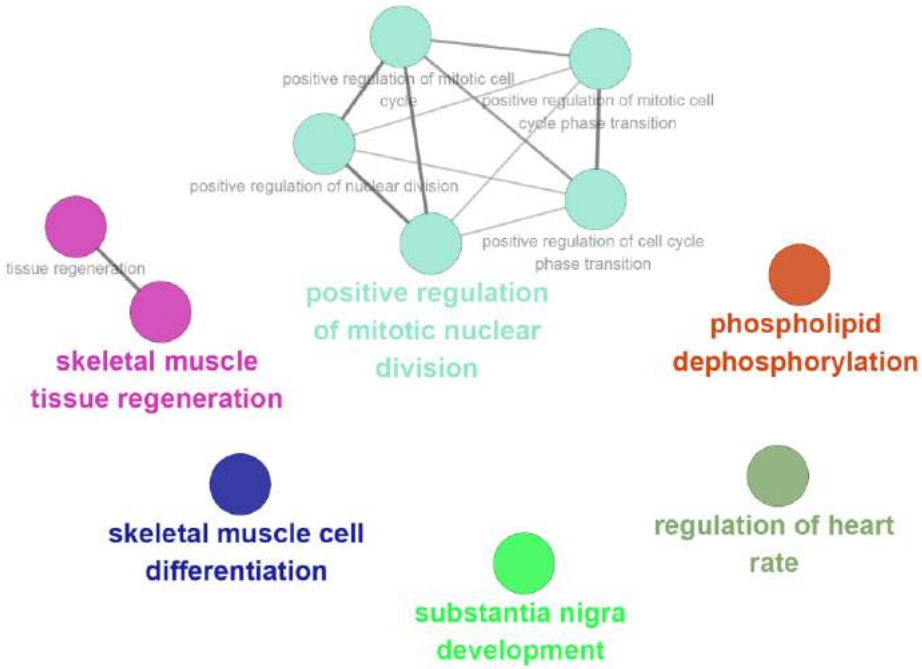



**Figure 6 - 3**. Enriched biological process Gene Ontology term network using 174 genes from unaligned read using ClueGO analysis.

focused on the muscle weight and quality of meat. Annexin A1(ANXA1), B-cell CLL/lymphoma 9 protein(BCL9) and SRY-related HMG-box(SOX15) genes were identified as components of the muscle related GO terms. Among these genes, ANXA1 is Ca2+-dependent phospholipid-binding proteins included in annexin family and it is preferentially located on the cytosolic face of the plasma membrane. It is the known candidate gene related with phenotypes of Hanwoo in previous study(Lee et al. 2013). This result indicates that remained genetic signatures of breeding in Hanwoo genome can be a structural variation and constructed pipeline can be effectively applied for various species. In addition, constructed pipeline is already used in previous studies of local breeds such as Korean Native Chicken(Kwak et al. 2014), Berkshire Pig(Jeong et al. 2015) and Blue egg chicken(Jeong et al. 2016) and the performance of constructed pipeline have been proved. I hope that constructed pipeline can provide more valuable variant information for local breeds and contribute to resolving of the limitations in current re-sequencing process.

## General Discussion

This study is mainly focused on the genome projects for unprecedented species and solving the limitation of previous studies. In chapter 2, I employed Pacbio long read sequencing system for constructing the complete genome of microbes. The Pacbio long read provide more effective and successful ways of genome assembly compared to Sanger sequencing and short-read NGS platform. Furthermore, this showed that changing sequencing paradigm could rapidly change the whole processes of previous analysis methods. I think the most important take home message from this study is that a bioinformatician should always be prepared to adapt to the new technologies and analysis algorithms.

Next study is the abalone genome project that was based on the Illumina sequencing platform with Pacbio long read. The goal of this study is on building the first draft genome of abalone, which is one of the most important economical animals in fishery industry. I successfully built the draft genome of abalone using two different sequencing platforms, and Pacbio long read showed its value in scaffolding process. Therefore, I think that long read sequencing system is one of the promised ways for the successful genome project. In addition, I expect that the draft genome presented here, which is the first genome to be sequenced in the family Haliotidae, will provide useful

genomic information for many researchers. Haliotis discus hannai is a cold-water abalone breed that has difficulties dealing with the change in their inhabitable latitude, which is due to global warming and the resulting increase in the rate of sudden perishing. Genomic information of abalone is essential information that can be used for genetic breeding to improve productivity and genetic engineering for the heat-resistant breed. It can also provide valuable information for future genomic studies because only a limited genome information about marine animals and mollusks is currently available. Evolutionary signatures recorded in abalone genome can be identified through future comparative genomic studies, and I expect my results will provide more insight into Haliotidae and marine mollusk evolution.

In chapter 4, constructed pipeline can be useful for building complete mitochondrial genome sequence compared to conventional Sanger sequencing. Sequencing cost is under 1/3 compared to primer walking method and constructed pipeline in this study can be carried out within one week, which is 1/4 compared to the previous method. To add, the success rate is much higher than the previous method. Therefore, mitochondrial genome assembly pipeline constructed in this study can resolve the limitations of the previous method, and I expect that this pipeline provides more opportunity for

researchers who are interested in the mitochondrial genome of unprecedented species.

Microsatellite marker build pipeline constructed in Chapter 5 can be employed as a useful tool for ecological studies. Under ordinary circumstances, genotyping microsatellite marker without reference genome is difficult to conduct; the reference genome is essential for designing primer sequences for genotyping. However, building reference genome needs a large amount of sequencing data, and this requires quite some research funds. If using our pipeline, researchers can conduct their population genetic studies using microsatellite markers with a reasonable price. I expect that constructed pipeline may support ecological studies of many researchers.

In the last chapter, I conducted unaligned read assembly for Hanwoo population. Constructed pipeline was successfully applied, and candidate structural variants that might be related with the breeding process of Hanwoo were identified. Network analysis on enriched GO term showed that identified genes are related to cell differentiation, muscle development and etc. Hanwoo has evolved to be meat type cattle over certain periods and the result of breeding program might be recorded in the Hanwoo genome. My results showed that unaligned read assembly is one of the effective methods for identifying genetic signatures in the local breed, and it supplements the limits

of re-sequencing studies based on the reference genome of the western breed. That said, constructed pipeline will provide more understanding of the relationship between genetic backgrounds and phenotypes.

Overall, I conducted genome projects for identifying recorded genetic signatures during evolution process and applied genome assembly method for various studies. I hope that my result can provide more opportunity to researchers that struggle with the analysis methods of unprecedented species, and give more insight into genome evolution of various living organisms.

## References


Adachi, K. and S.-I. Okumura (2012). "Determination of genome size of Haliotis discus hannai and H. diversicolor aquatilis (Haliotidae) and phylogenetic examination of this family." Fisheries science **78**(4): 849-852.

Albertin, C. B., O. Simakov, T. Mitros, Z. Y. Wang, J. R. Pungor, E. Edsinger-Gonzales, S. Brenner, C. W. Ragsdale and D. S. Rokhsar (2015). "The octopus genome and the evolution of cephalopod neural and morphological novelties." Nature **524**(7564): 220-224.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research **25**(17): 3389-3402.

An, H.-S., Y.-J. Jee, K.-S. Min, B.-L. Kim and S.-j. Han (2005). "Phylogenetic analysis of six species of Pacific abalone (Haliotidae) based on DNA sequences of 16s rRNA and cytochrome c oxidase subunit I mitochondrial genes." Marine biotechnology **7**(4): 373-380.

An, H., D. Jung, J. Lee and C.-B. Kim (2016). "The complete mitochondrial genome of Aplysia kurodai (Anaspidea: Aplysiidae)." Mitochondrial DNA part A **27**(2): 863-864.

Andrews, S. (2010). "FASTQC. A quality control tool for high throughput sequence data." URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Appeltans, W., P. Bouchet, G. Boxshall, K. Fauchald, D. Gordon, B. Hoeksema, G. Poore, R. Van Soest, S. Stöhr and T. Walter (2012). "World register of marine species." Accessed online: http://www. marinespecies. org (accessed on 28 February 2014).

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass and M. Kubal (2008). "The RAST Server: rapid annotations using subsystems technology." BMC genomics **9**(1): 75.

Bailly-Bechet, M., A. Haudry and E. Lerat (2014). "“One code to find them all”: a perl tool to conveniently parse RepeatMasker output files." Mobile DNA **5**(1): 1.

Baldwin, K. M. and F. Haddad (2001). "Invited review: effects of different activity and inactivity paradigms on myosin heavy chain gene expression in striated muscle." Journal of applied physiology **90**(1): 345-357.

Bao, Z. and S. R. Eddy (2002). "Automated de novo identification of repeat sequence families in sequenced genomes." Genome research **12**(8): 1269-1276.

Barreteau, H., A. Kovač, A. Boniface, M. Sova, S. Gobec and D. Blanot (2008). "Cytoplasmic steps of peptidoglycan biosynthesis." FEMS microbiology reviews **32**(2): 168-207.

Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov and E. S. Lander (2002). "ARACHNE: a whole-genome shotgun assembler." Genome research **12**(1): 177-189.

Benson, G. (1999). "Tandem repeats finder: a program to analyze DNA sequences." Nucleic acids research **27**(2): 573.

Bernt, M., A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf and P. F. Stadler (2013). "MITOS: Improved de novo metazoan mitochondrial genome annotation." Molecular phylogenetics and evolution **69**(2): 313-319.

Bindea, G., B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski and J. Galon (2009).

"ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks." Bioinformatics **25**(8): 1091-1093.

Blanco, E., G. Parra and R. Guigó (2007). "Using geneid to identify genes." Current protocols in bioinformatics: 4.3. 1-4.3. 28.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler and W. Pirovano (2011). "Scaffolding pre-assembled contigs using SSPACE." Bioinformatics **27**(4): 578-579.

Boge, T., M. Rémigy, S. Vaudaine, J. Tanguy, R. Bourdet-Sicard and S. Van Der Werf (2009). "A probiotic fermented dairy drink improves antibody response to influenza vaccination in the elderly in two randomised controlled trials." Vaccine **27**(41): 5677-5684.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics: btu170.

Bosch, A., G. Biesbroek, K. Trzcinski, E. Sanders and D. Bogaert (2013). "Viral and bacterial interactions in the upper respiratory tract." PLoS pathog **9**(1): e1003057.

Bottinelli, R., R. Betto, S. Schiaffino and C. Reggiani (1994). "Unloaded shortening velocity and myosin heavy chain and alkali light chain isoform composition in rat skeletal muscle fibres." The journal of physiology **478**(Pt 2): 341.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut and A. J. Drummond (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis." PLoS comput biol **10**(4): e1003537.

Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs and X. Huang (2008). "The

potential and challenges of nanopore sequencing." Nature biotechnology **26**(10): 1146-1153.

Bravo Portela, I., V. Martinez-Zorzano, I. Molist-Perez and P. Molist García (2012). "Ultrastructure and glycoconjugate pattern of the foot epithelium of the abalone Haliotis tuberculata (Linnaeus, 1758)(Gastropoda, Haliotidae)." The scientific world journal **2012**.

Bron, P. A., P. van Baarlen and M. Kleerebezem (2012). "Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa." Nature reviews microbiology **10**(1): 66-78.

Bryant, D. W., W.-K. Wong and T. C. Mockler (2009). "QSRA–a quality-value guided de novo short read assembler." BMC bioinformatics **10**(1): 69.

Burgess, S. A., M. L. Walker, H. Sakakibara, P. J. Knight and K. Oiwa (2003). "Dynein structure and power stroke." Nature **421**(6924): 715-718.

Butler, J., I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum and D. B. Jaffe (2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads." Genome research **18**(5): 810-820.

Carver, T., N. Thomson, A. Bleasby, M. Berriman and J. Parkhill (2009). "DNAPlotter: circular and linear interactive genome visualization." Bioinformatics **25**(1): 119-120.

Castoe, T. A., A. W. Poole, A. J. de Koning, K. L. Jones, D. F. Tomback, S. J. Oyler-McCance, J. A. Fike, S. L. Lance, J. W. Streicher and E. N. Smith (2012). "Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake." PLoS one **7**(2): e30953.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular biology and evolution **17**(4): 540-552.

Chénais, B., A. Caruso, S. Hiard and N. Casse (2012). "The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments." Gene **509**(1): 7-15.

Chaisson, M., P. Pevzner and H. Tang (2004). "Fragment assembly with short reads." Bioinformatics **20**(13): 2067-2074.

Chaisson, M. J., D. Brinza and P. A. Pevzner (2009). "De novo fragment assembly with short mate-paired reads: Does the read length matter?" Genome research **19**(2): 336-346.

Chaisson, M. J. and P. A. Pevzner (2008). "Short read fragment assembly of bacterial genomes." Genome research **18**(2): 324-330.

Chakraborty, R. and L. Jin (1993). "Determination of relatedness between individuals using DNA fingerprinting." Human biology: 875-895.

Charles, A. G., T.-Y. Han, Y. Y. Liu, N. Hansen, A. E. Giuliano and M. C. Cabot (2001). "Taxol-induced ceramide generation and apoptosis in human breast cancer cells." Cancer chemotherapy and pharmacology **47**(5): 444-450.

Chen, K., J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang and D. P. Locke (2009). "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation." Nature methods **6**(9): 677-681.

Chiancone, E. and P. Ceci (2010). "The multifaceted capacity of Dps proteins to combat bacterial stress conditions: detoxification of iron and

hydrogen peroxide and DNA binding." Biochimica et biophysica acta (BBA)-general subjects **1800**(8): 798-805.

Chin, C.-S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston and E. E. Eichler (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data." Nature methods **10**(6): 563-569.

Choi, S.-W., H.-N. Youn, W. Hong, J.-K. Park, S.-S. Yuk, J.-H. Kwon, J.-Y. Noh, J.-S. Kang, K.-J. Cho and J.-J. Ryu (2015). "Intranasal Administration Model for Evaluating Protection Against Influenza Virus in Mice." Journal of bacteriology and virology **45**(1): 44-50.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu and D. M. Ruden (2012). "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly **6**(2): 80-92.

Coleman, A. W. and V. D. Vacquier (2002). "Exploring the phylogenetic utility of ITS sequences for animals: a test case for abalone (Haliotis)." Journal of molecular evolution **54**(2): 246-257.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Consortium, U. (2011). "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic acids research: gkr981.

Cook, P. A. (2014). "The worldwide abalone industry." Modern economy **5**(13): 1181.

Coordinators, N. R. (2013). "Database resources of the national center for biotechnology information." Nucleic acids research **41**(Database issue): D8.

Davies, M. S. and S. Hawkins (1998). "Mucus from marine molluscs." Advances in marine biology **34**: 1-71.

De Bie, T., N. Cristianini, J. P. Demuth and M. W. Hahn (2006). "CAFE: a computational tool for the study of gene family evolution." Bioinformatics **22**(10): 1269-1271.

De Donato, M., S. O. Peters, S. E. Mitchell, T. Hussain and I. G. Imumorin (2013). "Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing." PLoS one **8**(5): e62137.

De Zoysa, M. (2013). "Nutritional Value, Bioactive Compounds, and Health-Promoting Properties of Abalone." Marine nutraceuticals: prospects and perspectives: 57.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Applied and environmental microbiology **72**(7): 5069-5072.

Dohm, J. C., C. Lottaz, T. Borodina and H. Himmelbauer (2007). "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing." Genome research **17**(11): 1697-1706.

Duboule, D. (2007). "The rise and fall of Hox gene clusters." Development **134**(14): 2549-2560.

Elliott, N. G. (2000). "Genetic improvement programmes in abalone: what is the future?" Aquaculture research **31**(1): 51-59.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler and S. E. Mitchell (2011). "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species." PLoS one **6**(5): e19379.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid and K. C. Worley (2012). "Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology." PloS one **7**(11): e47768.

Estes, J. A., D. R. Lindberg and C. Wray (2005). "Evolution of large body size in abalones (Haliotis): patterns and implications." Paleobiology **31**(4): 591-606.

Faircloth, B. C. (2008). "Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design." Molecular ecology resources **8**(1): 92-94.

Ferguson, N. M., D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley and D. S. Burke (2006). "Strategies for mitigating an influenza pandemic." Nature **442**(7101): 448-452.

Ferguson, N. M., C. Fraser, C. A. Donnelly, A. C. Ghani and R. M. Anderson (2004). "Public health risk from the avian H5N1 influenza epidemic." Science **304**(5673): 968.

Fritsch, M., T. Wollesen, A. de Oliveira and A. Wanninger (2015). "Unexpected co-linearity of Hox gene expression in an aculiferan mollusk." BMC evolutionary biology **15**(1): 1.

Glass, E. M., J. Wilkening, A. Wilke, D. Antonopoulos and F. Meyer (2010). "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes." Cold spring harbor protocols **2010**(1): pdb. prot5368.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea and S. Sykes (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proceedings of the national academy of sciences **108**(4): 1513-1518.

Gordon, H. R. and P. A. Cook (2004). "World abalone fisheries and aquaculture update: supply and market dynamics." Journal of shellfish Research **23**(4): 935-940.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury and Q. Zeng (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature biotechnology **29**(7): 644-652.

Gray, M. W., G. Burger and B. F. Lang (1999). "Mitochondrial evolution." Science **283**(5407): 1476-1481.

Guo, F., Z.-b. Huang, M.-q. Huang, J. Zhao and C.-h. Ke (2009). "Effects of small abalone, Haliotis diversicolor, pedal mucus on bacterial growth, attachment, biofilm formation and community structure." Aquaculture **293**(1): 35-41.

Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell and J. R. Wortman (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments." Genome biology **9**(1): R7.

Hahn, C., L. Bachmann and B. Chevreux (2013). "Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach." Nucleic acids research: gkt371.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic acids symposium series.

Hamer, P. A. and F. Victoria (2010). Understanding the ecological role of abalone in the reef ecosystem of victoria, Fisheries Victoria.

Handsaker, R. E., J. M. Korn, J. Nemesh and S. A. McCarroll (2011). "Discovery and genotyping of genome structural polymorphism by sequencing on a population scale." Nature genetics **43**(3): 269-276.

Harata, G., F. He, N. Hiruta, M. Kawase, A. Kubota, M. Hiramatsu and H. Yausi (2011). "Intranasally administered TMC0356 protects mice from H1N1 influenza virus infection by stimulating respiratory immune responses." World journal of microbiology and biotechnology **2**(27): 411-416.

Harridge, S., R. Bottinelli, M. Canepari, M. Pellegrino, C. Reggiani, M. Esbjörnsson and B. Saltin (1996). "Whole-muscle and single-fibre contractile properties and myosin heavy chain isoforms in humans." Pflügers archiv **432**(5): 913-920.

Hedges, S. B., J. Dudley and S. Kumar (2006). "TimeTree: a public knowledge-base of divergence times among organisms." Bioinformatics **22**(23): 2971-2972.

Herten, K., M. S. Hestand, J. R. Vermeesch and J. K. Van Houdt (2015). "GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments." BMC bioinformatics **16**(1): 1.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols **4**(1): 44-57.

Huang, X., J. Wang, S. Aluru, S.-P. Yang and L. Hillier (2003). "PCAP: a whole-genome assembly program." Genome research **13**(9): 2164-2170.

Hubisz, M. J., D. Falush, M. Stephens and J. K. Pritchard (2009). "Inferring weak population structure with the assistance of sample group information." Molecular ecology resources **9**(5): 1322-1332.

Hunt, M., N. De Silva, T. D. Otto, J. Parkhill, J. A. Keane and S. R. Harris (2015). "Circlator: automated circularization of genome assemblies using long sequencing reads." Genome biology **16**(1): 1.

Hutchinson, N., M. S. Davies, J. S. Ng and G. A. Williams (2007). "Trail following behaviour in relation to pedal mucus production in the intertidal gastropod Monodonta labio (Linnaeus)." Journal of experimental marine biology and ecology **349**(2): 313-322.

Hwang, H., T. Chen, R. G. Nines, H. C. Shin and G. D. Stoner (2006). "Photochemoprevention of UVB-induced skin carcinogenesis in SKH-1 mice by brown algae polyphenols." International journal of cancer **119**(12): 2742-2749.

Ibañez-Tallon, I., S. Gorokhova and N. Heintz (2002). "Loss of function of axonemal dynein Mdnah5 causes primary ciliary dyskinesia and hydrocephalus." Human molecular genetics **11**(6): 715-721.

Jang, M.-S., H.-Y. Park and K.-H. Nam (2012). "Changes in nutrient composition and fermentation properties of abalone mul-kimchi using dried pollack and licorice stock." Korean journal of food science and technology **44**(5): 613-620.

Jeck, W. R., J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl and C. D. Jones (2007). "Extending assembly of short DNA sequences to handle error." Bioinformatics **23**(21): 2942-2944.

Jehle, R. and J. Arntzen (2002). "Review: microsatellite markers in amphibian conservation genetics." Herpetological journal **12**: 1-9.

Jeong, H., K. Kim, K. Caetano-Anollés, H. Kim, B.-k. Kim, J.-K. Yi, J.-J. Ha, S. Cho and D. Y. Oh (2016). "Whole genome sequencing of Gyeongbuk Araucana, a newly developed blue-egg laying chicken breed, reveals its origin and genetic characteristics." Scientific reports **6**.

Jeong, H., K.-D. Song, M. Seo, K. Caetano-Anollés, J. Kim, W. Kwak, J.-d. Oh, E. Kim, D. K. Jeong and S. Cho (2015). "Exploring evidence of positive selection reveals genetic basis of meat quality traits in Berkshire pigs through whole genome sequencing." BMC genetics **16**(1): 1.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell and G. Nuka (2014). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.

Joseph, C., Y. Togawa and N. Shindo (2013). "Bacterial and viral infections associated with influenza." Influenza and other respiratory viruses **7**(s2): 105-113.

Jung, W.-K., S.-J. Heo, Y.-J. Jeon, C.-M. Lee, Y.-M. Park, H.-G. Byun, Y. H. Choi, S.-G. Park and I.-W. Choi (2009). "Inhibitory effects and molecular mechanism of dieckol isolated from marine brown alga on COX-2 and iNOS in microglial cells." Journal of agricultural and food chemistry **57**(10): 4439-4446.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and genome research **110**(1-4): 462-467.

Kajitani, R., K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu and H. Maruyama (2014). "Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads." Genome research **24**(8): 1384-1395.

Karagozlu, M. Z., J. Sung, J. Lee, W. Kwak and C.-B. Kim (2016). "Complete sequences of mitochondrial genome of Hypselodoris festiva (A. Adams, 1861)(Mollusca, Gastropoda, Nudibranchia)." Mitochondrial DNA part B: 1-2.

Karagozlu, M. Z., J. Sung, J. Lee, T. Kwon and C.-B. Kim (2016). "Complete mitochondrial genome sequences and phylogenetic relationship of Elysia ornata (Swainson, 1840)(Mollusca, Gastropoda, Heterobranchia, Sacoglossa)." Mitochondrial DNA part B: 1-2.

Kawamura, T., T. Saido, H. Takami and Y. Yamashita (1995). "Dietary value of benthic diatoms for the growth of post-larval abalone Haliotis discus hannai." Journal of experimental marine biology and ecology **194**(2): 189-199.

Kawamura, T. and H. Takami (1995). "Analysis of feeding and growth rate of newly metamorphosed abalone Haliotis discus hannai fed on four species of benthic diatom." Fisheries science **61**(2): 357-358.

Kechaou, N., F. Chain, J.-J. Gratadoux, S. Blugeon, N. Bertho, C. Chevalier, R. Le Goffic, S. Courau, P. Molimard and J. M. Chatel (2013). "Identification of one novel candidate probiotic Lactobacillus plantarum strain active against influenza virus infection in mice by a large-scale screening." Applied and environmental microbiology **79**(5): 1491-1499.

Kelley, D. R., M. C. Schatz and S. L. Salzberg (2010). "Quake: quality-aware detection and correction of sequencing errors." Genome biology **11**(11): R116.

Kidwell, M. G. (2002). "Transposable elements and the evolution of genome size in eukaryotes." Genetica **115**(1): 49-63.

Kim, K. M., S. Sung, G. Caetano-Anollés, J. Y. Han and H. Kim (2008). "An approach of orthology detection from homologous sequences under minimum evolution." Nucleic acids research **36**(17): e110-e110.

Kim, S.-K. and K. Chojnacka (2015). Marine algae extracts: processes, products, and applications, 2 Volume Set, John Wiley & Sons.

Kim, S. K., Y. Yong, S. H. Han, Y. S. Oh, M. H. Ko and M. Y. Oh (2000). "Phylogenetic relationship among Haliotis spp. distributed in Korea by the RAPD analysis." Korean journal of genetics **22**(1): 43-50.

Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." Journal of molecular evolution **16**(2): 111-120.

Koh, S.-M., H.-S. Kim, Y.-C. Cho, S.-G. Kang and J.-M. Kim (2009). "Preparation and physicochemical characteristics of abalone meat aged in Kochujang." Journal of the korean society of food science and nutrition **38**(6): 773-779.

Kozarewa, I., Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman and D. J. Turner (2009). "Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes." Nature methods **6**(4): 291-295.

Kroemer, G. and J. C. Reed (2000). "Mitochondrial control of cell death." Nature medicine **6**(5).

Kurinna, S. M., C. C. Tsao, A. F. Nica, T. Jiffar and P. P. Ruvolo (2004). "Ceramide promotes apoptosis in lung cancer-derived A549 cells by a mechanism involving c-Jun NH2-terminal kinase." Cancer research **64**(21): 7852-7856.

Kwak, W., K.-D. Song, J.-D. Oh, K.-N. Heo, J.-H. Lee, W. K. Lee, S. H. Yoon, H. Kim, S. Cho and H.-K. Lee (2014). "Uncovering Genomic Features and Maternal Origin of Korean Native Chicken by Whole Genome Sequencing." PloS one **9**(12): e114763.

Löytynoja, A. and N. Goldman (2005). "An algorithm for progressive multiple alignment of sequences with insertions." Proceedings of the

national academy of sciences of the United States of America **102**(30): 10557.

Löytynoja, A. and N. Goldman (2008). "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis." Science **320**(5883): 1632-1635.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Latuihamallo, M. and D. Apituley (2015). "Amino Acid and Fatty Acid of Abalone Haliotis Squamata Cultured in Different Aquaculture Systems." Procedia food science **3**: 174-181.

Lebeer, S., J. Vanderleyden and S. C. De Keersmaecker (2008). "Genes and molecules of lactobacilli supporting probiotic action." Microbiology and molecular biology reviews **72**(4): 728-764.

Lee, C.-G., H.-K. Kwon, J. H. Ryu, S. J. Kang, C.-R. Im, J. I. Kim and S.-H. Im (2010). "Abalone visceral extract inhibit tumor growth and metastasis by modulating Cox-2 levels and CD8+ T cell activity." BMC complementary and alternative medicine **10**(1): 60.

Lee, J.-S., s.-H. Won, S.-K. Kim, H. K. Lim and J. S. Lee (2014). "Classification and Description of Genus Nordotis (Gastropoda: Vestigastropoda) from Korea." The korean journal of malacology **30**(1): 79-86.

Lee, K.-T., W.-H. Chung, S.-Y. Lee, J.-W. Choi, J. Kim, D. Lim, S. Lee, G.-W. Jang, B. Kim and Y. H. Choy (2013). "Whole-genome resequencing of Hanwoo (Korean cattle) and insight into regions of homozygosity." BMC genomics **14**(1): 1.

LEE, S.-M., C. S. PARK and D. S. KIM (2001). "Effects of dietary herbs on growth and body composition of juvenile abalone, Haliotis discus hannai." Korean journal of fisheries and aquatic sciences **34**(5): 570-575.

Lee, Y.-H., T. Ota and V. D. Vacquier (1995). "Positive selection is a general phenomenon in the evolution of abalone sperm lysin." Molecular biology and evolution **12**(2): 231-238.

Leyer, G. J., S. Li, M. E. Mubasher, C. Reifer and A. C. Ouwehand (2009). "Probiotic effects on cold and influenza-like symptom incidence and duration in children." Pediatrics **124**(2): e172-e179.

Li, H. and R. Durbin (2011). "Inference of human population history from individual whole-genome sequences." Nature **475**(7357): 493-496.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, L., C. J. Stoeckert and D. S. Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome research **13**(9): 2178-2189.

Li, R., H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan and K. Kristiansen (2010). "De novo assembly of human genomes with massively parallel short read sequencing." Genome research **20**(2): 265-272.

Liévin-Le Moal, V. and A. L. Servin (2014). "Anti-infective activities of lactobacillus strains in the human intestinal microbiota: from probiotics to gastrointestinal anti-infectious biotherapeutic agents." Clinical microbiology reviews **27**(2): 167-199.

Lim, J., S. Park, J. Jeong, K. Park, K. Seo and J. Sung (2013). "Quality characteristics of kimchi fermented with abalone or sea tangle extracts." Journal of the korean society of food science and nutrition.

Lim, S. Y. (2014). "Cytotoxic and Antioxidant Activities of Abalone (Haliotis discus hannai) Extracts." Journal of life science **24**(7): 737-742.

Lin, A., R. Brunner, P. Chen, F. Talke and M. Meyers (2009). "Underwater adhesion of abalone: the role of van der Waals and capillary forces." Acta materialia **57**(14): 4178-4185.

Lin, A. and M. A. Meyers (2005). "Growth and structure in abalone shell." Materials science and engineering: A **390**(1): 27-41.

Lin, J., K. Okada, M. Raytchev, M. C. Smith and D. Nicastro (2014). "Structural mechanism of the dynein power stroke." Nature cell biology **16**(5): 479-485.

Lindberg, D. R. and W. F. Ponder (2001). "The influence of classification on the evolutionary interpretation of structure a re-evaluation of the evolution of the pallial cavity of gastropod molluscs." Organisms Diversity & Evolution **1**(4): 273-299.

Lischer, H. and L. Excoffier (2012). "PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs." Bioinformatics **28**(2): 298-299.

Liu, K. and S. V. Muse (2005). "PowerMarker: an integrated analysis environment for genetic marker analysis." Bioinformatics **21**(9): 2128-2129.

Liu, X., X. Liu, X. Guo, Q. Gao, H. Zhao and G. Zhang (2006). "A preliminary genetic linkage map of the Pacific abalone Haliotis discus hannai Ino." Marine Biotechnology **8**(4): 386-397.

Liu, Z.-H., L. Kang and J.-P. Wang (2014). "Basic and clinical research on the regulation of the intestinal barrier by Lactobacillus and its active protein components: a review with experience of one center." Molecular biology reports **41**(12): 8037-8046.

Liu, Z., Y. Ma, M. P. Moyer, P. Zhang, C. Shi and H. Qin (2012). "Involvement of the mannose receptor and p38 mitogen-activated protein kinase signaling pathway of the microdomain of the integral

membrane protein after enteropathogenic Escherichia coli infection." Infection and immunity **80**(4): 1343-1350.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan and Y. Liu (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." GigaScience **1**(1): 18.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan and Y. Liu (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." GigaScience **1**(1): 1-6.

Magoč, T. and S. L. Salzberg (2011). "FLASH: fast length adjustment of short reads to improve genome assemblies." Bioinformatics **27**(21): 2957-2963.

Mai, K., J. P. Mercer and J. Donlon (1995). "Comparative studies on the nutrition of two species of abalone, Haliotis tuberculata L. and Haliotis discus hannai Ino. IV. Optimum dietary protein level for growth." Aquaculture **136**(1): 165-180.

Maldonado-Barragán, A., J. L. Ruiz-Barba and R. Jiménez-Díaz (2009). "Knockout of three-component regulatory systems reveals that the apparently constitutive plantaricin-production phenotype shown by Lactobacillus plantarum on solid medium is regulated via quorum sensing." International journal of food microbiology **130**(1): 35-42.

Marçais, G. and C. Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." Bioinformatics **27**(6): 764-770.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen and Z. Chen (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Matsubara, T., M. Morita and A. Hayashi (1990). "Determination of the presence of ceramide aminoethylphosphonate and ceramide N-methylaminoethylphosphonate in marine animals by fast atom bombardment mass spectrometry." Biochimica et biophysica acta (BBA)-lipids and lipid metabolism **1042**(3): 280-286.

Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proceedings of the national academy of sciences **74**(2): 560-564.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel and M. Daly (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

Medvedev, P., M. Stanciu and M. Brudno (2009). "Computational methods for discovering structural variation with next-generation sequencing." Nature methods **6**: S13-S20.

Mendoza-Porras, O., N. A. Botwright, S. M. McWilliam, M. T. Cook, J. O. Harris, G. Wijffels and M. L. Colgrave (2014). "Exploiting genomic data to identify proteins involved in abalone reproduction." Journal of proteomics **108**: 337-353.

Menig, R., M. Meyers, M. Meyers and K. Vecchio (2000). "Quasi-static and dynamic mechanical response of Haliotis rufescens (abalone) shells." Acta materialia **48**(9): 2383-2398.

Metzker, M. L. (2009). "Sequencing technologies—the next generation." Nature reviews genetics **11**(1): 31-46.

Miller, J. R., A. L. Delcher, S. Koren, E. Venter, B. P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry and G. Sutton (2008). "Aggressive assembly of pyrosequencing reads with mates." Bioinformatics **24**(24): 2818-2824.

Miller, S. L. (1974). "Adaptive design of locomotion and foot form in prosobranch gastropods." Journal of experimental marine biology and ecology **14**(2): 99-156.

Moon, S. Y., H. S. Yoon, D. C. Seo and S. D. Choi (2006). "Growth Comparison of Juvenile Abalone." Haliotis discus hannai in different culture systems in the west coast of Kor, J Aquacult **19**: 242-246.

Morad, S. A. and M. C. Cabot (2013). "Ceramide-orchestrated signalling in cancer cells." Nature reviews cancer **13**(1): 51-65.

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert and K. A. Remington (2000). "A whole-genome assembly of Drosophila." Science **287**(5461): 2196-2204.

Neesen, J., R. Kirschner, M. Ochs, A. Schmiedl, B. Habermann, C. Mueller, A. F. Holstein, T. Nuesslein, I. Adham and W. Engel (2001). "Disruption of an inner arm dynein heavy chain gene results in asthenozoospermia and reduced ciliary beat frequency." Human molecular genetics **10**(11): 1117-1128.

O'Connell, J., O. Schulz-Trieglaff, E. Carlson, M. M. Hims, N. A. Gormley and A. J. Cox (2015). "NxTrim: optimized trimming of Illumina mate pair reads." Bioinformatics **31**(12): 2035-2037.

Ogretmen, B. and Y. A. Hannun (2004). "Biologically active sphingolipids in cancer pathogenesis and treatment." Nature reviews cancer **4**(8): 604-616.

Pádua, D., E. Rocha, D. Gargiulo and A. Ramos (2015). "Bioactive compounds from brown seaweeds: Phloroglucinol, fucoxanthin and fucoidan as promising therapeutic agents against breast cancer." Phytochemistry letters **14**: 91-98.

Panhuis, T. M., N. L. Clark and W. J. Swanson (2006). "Rapid evolution of reproductive proteins in abalone and Drosophila." Philosophical transactions of the royal society of london B: biological sciences **361**(1466): 261-268.

Park, J. Y., Y.-R. An, N. Kanda, C.-M. An, H. S. An, J.-H. Kang, E. M. Kim, D.-H. An, H. Jung and M. Joung (2015). "Cetaceans evolution: insights from the genome sequences of common minke whales." BMC genomics **16**(1): 1.

Park, M.-K., V. Ngo, Y.-M. Kwon, Y.-T. Lee, S. Yoo, Y.-H. Cho, S.-M. Hong, H. S. Hwang, E.-J. Ko and Y.-J. Jung (2013). "Lactobacillus plantarum DK119 as a probiotic confers protection against influenza virus by modulating innate immunity." PloS one **8**(10): e75368.

Parra, G., K. Bradnam and I. Korf (2007). "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes." Bioinformatics **23**(9): 1061-1067.

Peng, Y., H. C. Leung, S.-M. Yiu and F. Y. Chin (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428.

Pennarun, G., E. Escudier, C. Chapelin, A.-M. Bridoux, V. Cacheux, G. Roger, A. Clément, M. Goossens, S. Amselem and B. Duriez (1999). "Loss-of-function mutations in a human gene related to Chlamydomonas reinhardtii dynein IC78 result in primary ciliary dyskinesia." The american journal of human genetics **65**(6): 1508-1519.

Peterson, K. J., J. A. Cotton, J. G. Gehling and D. Pisani (2008). "The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records." Philosophical transactions of the royal society of london B: biological sciences **363**(1496): 1435-1443.

Pevzner, P. A., H. Tang and M. S. Waterman (2001). "An Eulerian path approach to DNA fragment assembly." Proceedings of the National Academy of Sciences **98**(17): 9748-9753.

Posada, D. and K. A. Crandall (1998). "Modeltest: testing the model of DNA substitution." Bioinformatics **14**(9): 817-818.

Price, A. L., N. C. Jones and P. A. Pevzner (2005). "De novo identification of repeat families in large genomes." Bioinformatics **21**(suppl 1): i351-i358.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic acids research **35**(suppl 1): D61-D65.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker and M. J. Daly (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The american journal of human genetics **81**(3): 559-575.

Ragg, N. L. and H. H. Taylor (2006). "Oxygen uptake, diffusion limitation, and diffusing capacity of the bipectinate gills of the abalone, Haliotis iris (Mollusca: Prosobranchia)." Comparative biochemistry and physiology part A: molecular & integrative physiology **143**(3): 299-306.

Ragg, N. L. C. (2003). Respiratory circulation in the abalone Haliotis iris, University of Canterbury.

Reynolds, C. P., B. J. Maurer and R. N. Kolesnick (2004). "Ceramide synthesis and metabolism as a target for cancer therapy." Cancer letters **206**(2): 169-180.

Richter, M. and R. Rosselló-Móra (2009). "Shifting the genomic gold standard for the prokaryotic species definition." Proceedings of the national academy of sciences **106**(45): 19126-19131.

Rizzardini, G., D. Eskesen, P. C. Calder, A. Capetti, L. Jespersen and M. Clerici (2012). "Evaluation of the immune benefits of two probiotic strains Bifidobacterium animalis ssp. lactis, BB-12® and Lactobacillus paracasei ssp. paracasei, L. casei 431® in an influenza vaccination model: a randomised, double-blind, placebo-controlled study." British journal of nutrition **107**(06): 876-884.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew and M. Edwards (2011). "An integrated semiconductor device enabling non-optical genome sequencing." Nature **475**(7356): 348-352.

Rothberg, J. M. and J. H. Leamon (2008). "The development and impact of 454 sequencing." Nature biotechnology **26**(10): 1117-1124.

Rozen, S. and H. Skaletsky (1999). "Primer3 on the WWW for general users and for biologist programmers." Bioinformatics methods and protocols: 365-386.

Ruan, J., H. Li, Z. Chen, A. Coghlan, L. J. M. Coin, Y. Guo, J.-K. Heriche, Y. Hu, K. Kristiansen and R. Li (2008). "TreeFam: 2008 update." Nucleic acids research **36**(suppl 1): D735-D740.

Sander, B., A. Nizam, L. P. Garrison, M. J. Postma, M. E. Halloran and I. M. Longini (2009). "Economic evaluation of influenza pandemic mitigation strategies in the United States using a stochastic microsimulation transmission model." Value in health **12**(2): 226-233.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the national academy of sciences **74**(12): 5463-5467.

Sekino, M. and M. Hara (2007). "Linkage maps for the Pacific abalone (genus Haliotis) based on microsatellite DNA markers." Genetics **175**(2): 945-958.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research **13**(11): 2498-2504.

Shepherd, S., M. J. Tegner and S. Guzmán del Próo (1992). "Abalone of the world: biology, fisheries and culture. Proceedings."

Shin, E.-S., K.-A. Lee, H.-K. Lee, K.-B.-W.-R. Kim, M.-J. Kim, M.-W. Byun, J.-W. Lee, J.-H. Kim, D.-H. Ahn and E.-S. Lyu (2008). "Effect of grain size and added water on quality characteristics of abalone porridge." Journal of the korean society of food science and nutrition **37**(2): 245-250.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics **31**(19): 3210-3212.

Simakov, O., F. Marletaz, S.-J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D.-H. Kuo, T. Larsson, J. Lv and D. Arendt (2013). "Insights into bilaterian evolution from three spiralian genomes." Nature **493**(7433): 526-531.

Siva, N. (2008). "1000 Genomes project." Nature biotechnology **26**(3): 256-256.

Slater, G. S. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC bioinformatics **6**(1): 31.

Smith, G. J., D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani and S. Bhatt (2009).

"Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic." Nature **459**(7250): 1122-1125.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent and L. E. Hood (1986). "Fluorescence detection in automated DNA sequence analysis."

Solovyev, V., P. Kosarev, I. Seledsov and D. Vorobyev (2006). "Automatic annotation of eukaryotic genes, pseudogenes and promoters." Genome biology **7**(Suppl 1): S10.

Stanke, M., M. Diekhans, R. Baertsch and D. Haussler (2008). "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics **24**(5): 637-644.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern (2006). "AUGUSTUS: ab initio prediction of alternative transcripts." Nucleic acids research **34**(suppl 2): W435-W439.

Stieb, M. and B. Schink (1984). "A new 3-hydroxybutyrate fermenting anaerobe, Ilyobacter polytropus, gen. nov. sp. nov., possessing various fermentation pathways." Archives of microbiology **140**(2-3): 139-146.

Stockdale, F. E. and J. B. Miller (1987). "The cellular basis of myosin heavy chain isoform expression during development of avian skeletal muscles." Developmental biology **123**(1): 1-9.

Sturme, M. H., C. Francke, R. J. Siezen, W. M. de Vos and M. Kleerebezem (2007). "Making sense of quorum sensing in lactobacilli: a special focus on Lactobacillus plantarum WCFS1." Microbiology **153**(12): 3939-3947.

Suleria, H. R., P. Masci, G. Gobe and S. Osborne (2015). "Therapeutic Potential of Abalone and Status of Bioactive Molecules: A Comprehensive Review." Critical reviews in food science and nutrition(just-accepted): 00-00.

Sung, J.-M., J. Lee, S.-G. Kim, M. Z. Karagozlu and C.-B. Kim (2016). "Analysis of complete mitochondrial genome of Ocypode cordimanus (Latreille, 1818)(Decapoda, Ocypodidae)." Mitochondrial DNA part B **1**(1): 363-364.

Sung, J.-M., J. Lee, S.-G. Kim, M. Z. Karagozlu and C.-B. Kim (2016). "Complete mitochondrial genome of Leptodius sanguineus (Decapoda, Xanthidae)." Mitochondrial DNA part B **1**(1): 500-501.

Sung, J.-M., J. Lee, S.-K. Kim, M. Z. Karagozlu and C.-B. Kim (2016). "The complete mitochondrial genome of Grapsus tenuicrustatus (Herbst, 1783)(Decapoda, Grapsidae)." Mitochondrial DNA part B **1**(1): 441-442.

Suo, C., Y. Yin, X. Wang, X. Lou, D. Song, X. Wang and Q. Gu (2012). "Effects of lactobacillus plantarum ZJ316 on pig growth and pork quality." BMC veterinary research **8**(1): 89.

Swanson, W. J. and V. D. Vacquier (1998). "Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein." Science **281**(5377): 710-712.

Sweeney, H. L., M. Kushmerick, K. Mabuchi, F. Sreter and J. Gergely (1988). "Myosin alkali light chain and heavy chain variations correlate with altered shortening velocity of isolated skeletal muscle fibers." Journal of biological chemistry **263**(18): 9034-9039.

Swofford, D. L. (2003). "{PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.}."

Talavera, G. and J. Castresana (2007). "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments." Systematic biology **56**(4): 564-577.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar (2013). "MEGA6: molecular evolutionary genetics analysis version 6.0." Molecular biology and evolution: mst197.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar (2013). "MEGA6: molecular evolutionary genetics analysis version 6.0." Molecular biology and evolution **30**(12): 2725-2729.

Tarailo-Graovac, M. and N. Chen (2009). "Using RepeatMasker to identify repetitive elements in genomic sequences." Current protocols in Bioinformatics: 4.10. 11-14.10. 14.

Taylor, H. and N. Ragg (2005). "The role of body surfaces and ventilation in gas exchange of the abalone, Haliotis iris." Journal of comparative physiology B **175**(7): 463-478.

Taylor, S. (2011). Marine medicinal foods: implications and applications, macro and microalgae, Academic Press.

Thomas, N. V. and S.-K. Kim (2011). "Potential pharmacological applications of polyphenolic derivatives from marine brown algae." Environmental toxicology and pharmacology **32**(3): 325-335.

Toshiko, M. and H. Akira (1982). "Structural studies on glycolipid of shellfish: IV. A novel pentaglycosylceramide from abalone, Haliotis japonica." Biochimica et biophysica acta (BBA)-lipids and lipid metabolism **711**(3): 551-553.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature protocols **7**(3): 562-578.

Travers, K. J., C.-S. Chin, D. R. Rank, J. S. Eid and S. W. Turner (2010). "A flexible and efficient template format for circular consensus sequencing and SNP detection." Nucleic acids research **38**(15): e159-e159.

Viana, M. T., L. M. López, Z. García-Esquivel and E. Mendez (1996). "The use of silage made from fish and abalone viscera as an ingredient in abalone feed." Aquaculture **140**(1): 87-98.

Vinther, J., E. A. Sperling, D. E. Briggs and K. J. Peterson (2012). "A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors." Proceedings of the royal society of london B: biological sciences **279**(1732): 1259-1268.

Voskoboynik, A., N. F. Neff, D. Sahoo, A. M. Newman, D. Pushkarev, W. Koh, B. Passarelli, H. C. Fan, G. L. Mantalas and K. J. Palmeri (2013). "The genome sequence of the colonial chordate, Botryllus schlosseri." Elife **2**.

Warren, R. L., G. G. Sutton, S. J. Jones and R. A. Holt (2007). "Assembling millions of short DNA sequences using SSAKE." Bioinformatics **23**(4): 500-501.

Whalen, R. G., S. M. Sell and G. S. Butler-Browne (1981). "Three myosin heavy-chain isozymes appear sequentially in rat muscle development." Nature **292**: 805-809.

Wu, S., Z. Zhu, L. Fu, B. Niu and W. Li (2011). "WebMGA: a customizable web server for fast metagenomic sequence analysis." BMC genomics **12**(1): 444.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular biology and evolution **24**(8): 1586-1591.

Yang, Z., W. J. Swanson and V. D. Vacquier (2000). "Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites." Molecular biology and evolution **17**(10): 1446-1455.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler and Z. Ning (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." Bioinformatics **25**(21): 2865-2871.

Yonge, C. "I947 The pallial organs in the aspidobranch Gastropoda and their evolution throughout the Mollusca." Philosophical transactions of the royal society B **232**: 443-518.

Yoo, M.-J. and H.-J. Chung (2007). "Optimal manufacturing condition and quality properties of the drinking extract of disk abalone." Journal of the korean society of food culture **22**(6): 827-832.

Youn, H.-N., D.-H. Lee, Y.-N. Lee, J.-K. Park, S.-S. Yuk, S.-Y. Yang, H.-J. Lee, S.-H. Woo, H.-M. Kim and J.-B. Lee (2012). "Intranasal administration of live Lactobacillus species facilitates protection against influenza virus infection in mice." Antiviral research **93**(1): 138-143.

Zelaya, H., A. Tada, M. G. Vizoso-Pinto, S. Salva, P. Kanmani, G. Agüero, S. Alvarez, H. Kitazawa and J. Villena (2015). "Nasal priming with immunobiotic Lactobacillus rhamnosus modulates inflammation–coagulation interactions and reduces influenza virus-associated pulmonary damage." Inflammation research **64**(8): 589-602.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821-829.

Zhang, G., X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang, L. Zhang, X. Wang and H. Qi (2012). "The oyster genome reveals stress adaptation and complexity of shell formation." Nature **490**(7418): 49-54.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie and B. S. Weir (2012). "A high-performance computing toolset for relatedness and principal component analysis of SNP data." Bioinformatics **28**(24): 3326-3328.

# 새로운 종에 대한 유전체 조립전략 수립

**곽우리**

**생물정보협동과정 생물정보학전공**

**서울대학교 대학원 자연과학대학**

염기서열 분석 기술의 개발은 기존에 연구된 적이 없는 유전체 재구축의 가능성을 높여주었다. 이런 참조유전체가 없는 종의 유전체의 재구축은 해당 종의 유전적 특성을 파악하고 진화적 역사를 파악하는데 매우 효과적이다. 그러나 최근 Genome Project 가 기존에 비해 좀 더 합리적인 가격으로 수행이 가능 해졌지만 여전히 Genome Project 를 수행하기 위해서는 상당한 양의 연구비가 필요하다. 그래서 모델동물이나 기존에 연구된 동물이 아닌 동물을 연구하는 연구자는 해당 동물의 유전체를 연구하는데 어려움을 겪고 있다. 이런 어려움을 해결하기 위해서 본문의 연구는 유전체 조립을 이용하여 참조유전체가 없는 종의 유전체를 재구축함과 동시에 기존 연구방법의 한계점을 보완할 수 있는 실용적 적용방법을 담고 있다. 이를 위해 고유한 특징을 가진 다양한 프로그램을 이용하여 각 연구 목적에 맞는 맞춤 파이프라인을 구축하였다.

챕터 1 에서는 기존의 차세대 염기서열 분석의 기본 배경지식과 함께 유전체 조립을 정리하였다. 현재 활용되고 있는 염기서열 분석 기술의 특징과 유전체조립 알고리즘 및 대표 프로그램들을 정리하였다.

챕터 2 에서는 미생물 Genome Project 를 수행하였다. Pacbio RS II 데이터를 활용하여 미생물의 Complete Genome 을 재구축 하였다. 이를 기반으로 비교유전체 분석을 수행하여 해당 미생물의 면역기능 강화 효과를 설명할 수 있는 유전적 배경을 확인하였다.

챕터 3 에서는 다양한 종류의 염기서열 분석 자료를 이용하여 전복의 Genome Project 를 수행하였다. 이는 전복류에서 수행된 첫번째 Genome Project 이며 이를 통해 유전체 비교분석을 기반으로 전복의 유전체에 기록된 유전적 특징을 확인하였다. 또한 전복의 유전체뿐만 아니라 전복 내장의 Metagenome 및 Metatranscriptome 분석을 통해 아직 알려지지 않은 전복 내장추출물의 항암효과에 대한 기원을 파악하고자 하였다.

챕터 4 에서는 선행연구가 없는 종에 대한 효과적인 미토콘드리아 유전체 재구축 파이프라인을 구축하였다. 이 파이프라인은 각기 다른 목적의 연구에서 활용된 여러

과정들이 혼합되어 있다. 구축된 파이프라인은 *M.tuberosa* 의 실제 데이터를 이용하여 검증되었다.

챕터 5 에서는 유전체조립을 이용하여 참조유전체가 없는 종의 초위성체 마커 개발을 위한 파이프라인를 구축하였다. 구축된 파이프라인은 천잠 유전체 데이터에 적용하였으며 집단유전학 연구에 활용될 수 있는 초위성체 마커 개발에 사용될 수 있음을 확인하였다.

마지막 챕터는 Unaligned read assembly 에 관련된 내용이다. Unaligned read assembly 는 참조유전체를 기반으로 하는 Re-sequencing study 의 한계를 보완할 수 있는 가장 효과적인 방법 중 하나다. 이를 한우 집단유전체 데이터에 적용하여 기존의 Re-sequencing 방법에서 확인하지 못했던 표현형 관련 유전체 구조변이를 확인 하였다.

나는 이 연구와 구축된 관련 파이프라인들이 참조유전체 서열이 없는 종의 유전체에 대한 이해와 이 분야의 기존 연구방법들의 한계를 개선하는데 기여할 것으로 기대한다.

**주요어** : 유전체 조립, 진화분석, 미토콘드리아, 초위성체, Unaligned read

**학 번** : 2014 - 30098